An Imputation-Estimation Algorithm Using Time-Varying Auxiliary Covariates for a Longitudinal Model When Outcome is Missing by Design

By Marinella Gracia Montealegre Temprosa

Bachelors in Statistics & Ancient Greek and Latin, May 1993, Rutgers University

Masters in Statistics, December 1996, George Washington University

A Dissertation submitted to

The Faculty of
Columbian College of Arts and Sciences
of The George Washington University in partial satisfaction
of the requirements for the degree of Doctor of Philosophy

August 31, 2012

Dissertation directed by

John M. Lachin III

Professor of Biostatistics and of Epidemiology, and Statistics

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

The Columbian College of Arts and Sciences of The George Washington University certifies that Marinella Temprosa has passed the Final Examination for the degree of Doctor of Philosophy as of July 24, 2012. This is the final and approved form of the dissertation.

An Imputation-Estimation Algorithm Using Time-Varying Auxiliary Covariates for a Longitudinal Model When Outcome is Missing by Design

Marinella Gracia Montealegre Temprosa

Dissertation Research Committee:

John M. Lachin III, Professor of Biostatistics and Epidemiology, and of Statistics, Dissertation Director

Michael Larsen, Associate Professor of Statistics, Committee Member

Qing Pan, Assistant Professor of Statistics, Committee Member

Dedication

I dedicate this dissertation to my mother Luau Montealegre, and grandparents Jose and Trinidad Montealegre who worked hard to provide me the opportunities and values to believe that anything is possible.

Acknowledgments

This dissertation was completed through the help, encouragement and love from my family, colleagues at the Biostatistics Center, and collaborators from the Diabetes Prevention Program Research Group. I am particularly indebted to my dissertation director, Professor John Lachin whose encouragement and support lasted for many years while I was working on my research. His invaluable contributions to the understanding of Type 1 and Type 2 diabetes is of great importance to public health and trailblazing for a biostatistician like me. I am deeply indebted to my tireless readers, Professors Qing Pan and Mike Larsen whose very careful review made this dissertation much improved and my thinking more clear. I cannot thank you enough from the bottom of my heart in helping me become a better statistician. I am blessed and grateful to have Professors Efstathia Bura, Nancy Cook, and Dante Verme as my dissertation committee since they are instrumental in my chosen doctoral pursuits. At a pivotal time when I was considering whether to pursue a doctoral degree in Biostatistics, I took Stat 242 Nonparametric Regression with Professor Bura. I was so inspired from the topics I learned from the course that I decided to pursue the doctoral degree. Professor Dante Verme is also instrumental as he was my first point of contact after admission to the doctoral program in Biostatistics. I am deeply honored to have Professor Nancy Cook on my committee since her papers from 1997 and 2006 provided me with the spark and direction for this research including the framework and skeleton for the algorithm. I am also grateful to Lisa Mele and Nisha Grover, really great friends and colleagues. Lisa read my proposal and dissertation even if she is busy pursuing her degree in architecture. Nisha not only

helped me with the printing and binding of this dissertation, she has become my family over the years.

I would also like to thank the founders of the George Washington Biostatistics Center for providing me a place to grow and learn from great mentors (Sarah Fowler, John Lachin, and Naji Younes) and dedicated colleagues. Sarah Fowler has given me many amazing opportunities to grow as an investigator and I have learned from her how to become a great leader and effective collaborator. I could not have been luckier to find a great teacher in Naji on all things technical and fun, not only in programming for statistical analyses but also study websites, data capture systems, and phone randomization systems. With great intellectual humility, I liken myself to one of the dwarfs on the shoulders of these giants (Latin: nani gigantum humeris insidentes) as John of Salisbury wrote in his Metalogicon (1159) *"... we are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness of sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size..."*.

Most importantly, a big thank you to my family (Robert, Jay and Nico Campanell) who supported me throughout this long process with great patience, love and a lot of sacrifice. The Biostatistics Center not only provided me a career but it also helped me in finding my partner for life, Rob. I look forward to spending more of my free time with my boys. Finally, my father Alfredo Temprosa whose love and care I felt from afar.

DPP Acknowledgements. The Diabetes Prevention Program (DPP) was conducted by the DPP Research Group and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the General Clinical Research Center Program, the National Institute of Child Health and Human Development (NICHD), the National Institute on Aging (NIA), the Office of Research on Women's Health, the Office of Research on Minority Health, the Centers for Disease Control and Prevention (CDC), and the American Diabetes Association. The data were supplied by the NIDDK Central Repositories. This analyses was not prepared under the auspices of the DPP and does not

represent analyses or conclusions of the DPP Research Group, the NIDDK Central Repositories, or the NIH.

Abstract

An Imputation-Estimation Algorithm Using Time-Varying Auxiliary Covariates for a Longitudinal Model When Outcome is Missing by Design

In long term clinical trials, occurrence of missing data is an area of concern especially if the rate at which data are missing depends on the treatment group. Typically, some effort is spent on trying to identify the reasons the data are missing so that appropriate assumptions and analytic approaches can be properly applied. When data are missing by design, certain measurements are discontinued after meeting an endpoint, possibly due to ethical or financial constraints. Subjects who reach the absorbing barrier may stop data collection on some variables but may subsequent time-varying covariates available from continued follow-up. In this dissertation, we developed an Imputation-Estimation algorithm under an auxiliary missing at random assumption to assess whether the additional information from the time varying covariates can be used to improve estimation. Quality of estimates is evaluated in terms of bias, variance and coverage for the estimates of the parameters of interest. We contrast this method to other missing data approaches such as multiple imputation and available case analysis.

We illustrate this method using data from the Diabetes Prevention Program (DPP). The DPP was a diabetes prevention study that showed reductions of 58% and 31% in diabetes risk using intensive lifestyle or metformin interventions compared to placebo. According to the DPP protocol, the oral glucose tolerance test is discontinued after diabetes diagnosis. Because of the significant reduction in diabetes incidence by the metformin and lifestyle interventions, the rates of missing IGR and CIR are different among the

treatment groups. This differential discontinuation among treatment groups results in informative monotone missing assessments of 30 minute glucose and insulin values. These 30 minute values are used to calculate surrogate measures of insulin secretion such as Insulin Glucose Ratio (IGR = (30-min insulin - fasting insulin)/(30-min glucose - fasting glucose)). Fasting blood glucose is collected at all time points and is associated with 30-minute glucose. The imputation estimation algorithm is applied to estimate the mean 30 minute blood glucose utilizing auxiliary information from the fasting blood glucose. In this example, fasting glucose is also the source of the discontinuation since diabetes diagnosis is based on the fasting glucose and 2 hour values during the OGTT. Because of the strong dependence between the fasting and 30 minute glucose measured at the same visit, the resulting estimates from the IE algorithm using the complete vector were similar to multiple imputation. Because the Placebo group experienced higher rates of diabetes incidence, the difference between available case analysis and the regression based imputations were greater than in the lifestyle group.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

In studies of chronic illnesses requiring long-term follow-up, investigations may include associations of a disease state and its effect on continuous or categorical outcomes. Missing data are relatively common in research studies and should be addressed appropriately because failure to do so may lead to invalid conclusions. When considering options for the handling of missing data one should consider the rate at which data are missing, the relationship among the missing and observed data, and the potential reasons that the data are missing. In this dissertation, we propose to study missing data arising from discontinuation of the measurement after reaching an absorbing disease state which was pre-specified in the study protocol (i.e., missing by design). Herein, we evaluate whether we could employ information gained from these correlated measures taken after the discontinuation of other measurements to recover lost information.

We explore the problem of missing data in the setting of a randomized clinical trial. For this research, we consider a longitudinal study with subjects having a series of scheduled assessments that will result in correlated data within a subject. The vector of outcome measurements and the time varying correlated measurements are assumed to follow a multivariate normal distribution under a linear mixed model framework. The

Imputation-Estimation approach allows us to model the conditional expectation of the outcome given the observed outcome measures, and correlated measurements before and after the endpoint. The planned missing data in our motivating example is described in the next section. The goal is to use auxiliary variables to make valid inferences about the outcome of interest.

1.2 Motivating Example

The evolution of Type 2 diabetes is marked by resistance of muscles to insulin absorption as well as reduced insulin secretion in the pancreas. This regulatory system is usually referred to as beta cell function. Thus, the monitoring of insulin sensitivity and secretion is important in the characterization of subjects in diabetes prevention studies. However, the "gold standard" measurement of insulin and sensitivity is difficult in large clinical studies due to the effort required to measure these parameters by euglycemic clamp (Wallace and Matthews, 2002). In large epidemiological studies, homeostasis based insulin resistance ($HOMA\ IR = (fasting\ insulin \times fasting\ glucose)/22.5$) (Matthews et al., 1985) and insulin to glucose ratio
($IGR = (30'\ insulin - fasting\ insulin)/(30'\ glucose\ -\ fasting\ glucose)$ (Phillips et al., 1994) are often used as surrogates for insulin sensitivity and secretion and are calculated from Oral Glucose Tolerance Tests (OGTT). The OGTT entails collection of samples at 2 time points (30 minutes and 2 hours) after administration of a 75-gram glucose load to a subject in a fasting state. The painstaking collection of OGTT samples makes this test ethically and fiscally unjustifiable after the diagnosis of diabetes in prevention studies. However, there may be covariates that are still measured that may give insight into the beta cell function. In this dissertation, we propose to explore methods for describing the insulin secretion over time so treatment effects may be assessed post diabetes diagnosis. The Diabetes Prevention Program is a randomized clinical trial that was designed to compare the effectiveness of intensive lifestyle or metformin therapy versus placebo on the

development of diabetes (DPP Research Group, 1999). The study recruited high risk volunteers and randomized 3234 participants to one of 3 treatment groups: Placebo, Metformin or Intesive Lifestyle. The diagnosis of diabetes was based on a confirmed biannual collection of fasting glucose and annual oral glucose tolerance tests using American Diabetes Association (1997) and World Health Organization (Alberti and Zimmet, 1998) criteria (fasting glucose $\geq$ 126 mg/dl or 2 hour glucose $\geq$ 200 mg/dl). According to the DPP protocol, glucose and insulin were measured at the fasting state semi-annually and during oral glucose tolerance test (OGTT) annually for the diagnosis of diabetes. The OGTT included collections at 30 minute for insulin and glucose and 2 hour for glucose only after a 75 gram glucose challenge but were discontinued after the participant was diagnosed with diabetes. Semi-annual follow-up visits continued for these subjects after diabetes diagnosis and included collection of correlated measures that may continue to provide an insight to the beta cell function. For example, fasting glucose and HbA1c were continued to be measured every 6 months and fasting insulin measured annually. We illustrate this protocol feature by example. A subject followed for 6 years and confirmed diabetic based on their fasting glucose on the 3^{rd} visit had missing 30 minute glucose values starting from the 4^{th} visit according to protocol (Table 1.1). The subject continued their visits for another 3 years when HbA1c and fasting glucose were measured and both correlate with 30 minute glucose.

Table 1.1: Sample data vector for subject diagnosed at 3rd visit

Measurement	Time 0	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Diabetes	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Fasting glucose	111	94	107	**131**	**109**	**101**	**99**
30 minute glucose	144	136	149	**168**	.	.	.
HbA1c	7.0	6.8	6.7	7.9	6.3	6.2	6.3

Figure 1.1 on the next page shows the mean glucose levels during baseline and year 1 OGTT according to treatment group (DPP Research Group, 2005). In this figure, the the glucose response to the 75 gram glucose challenge did not differ at baseline. At year 1, the

glucose responses differed among the 3 treatment groups with the greatest decreases observed in the 30 and 120 minute responses among the lifestyle group. In this analyses, a small number of participants diagnosed with diabetes at 6 months no longer contributed OGTT data at year 1 and progressively missing values among diabetics precluded analyses of further years. Our goal in this dissertation is to see whether we can an use time varying auxiliary covariates to get an insight into the glucose response in future years.

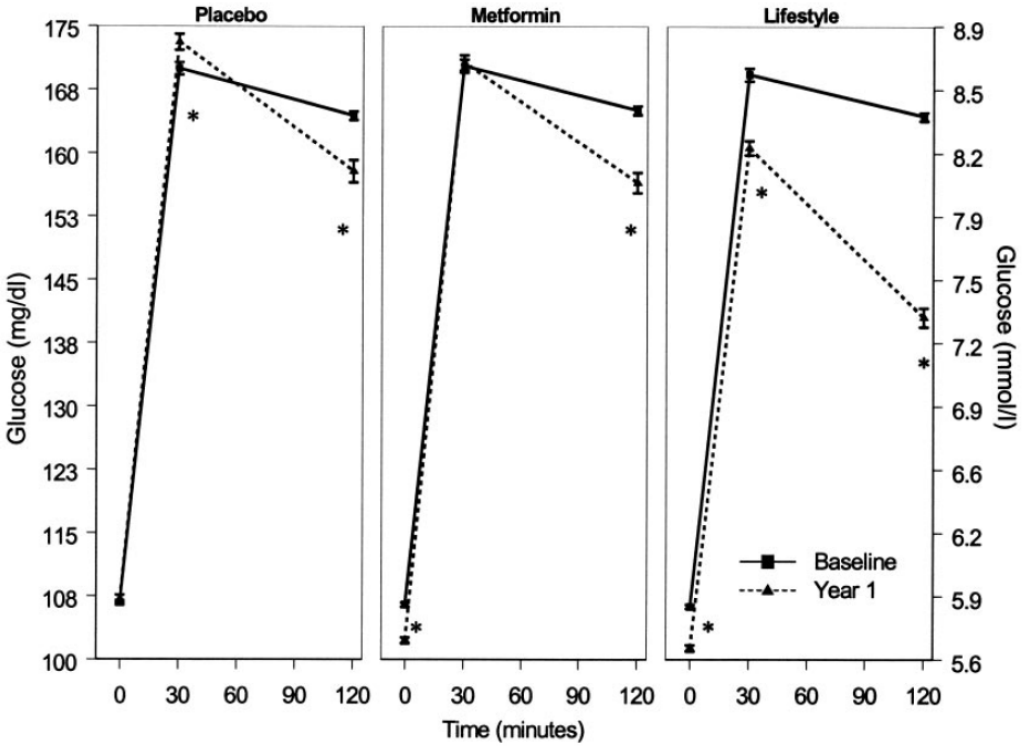


Figure 1.1: Mean plasma glucose (and standard error) by treatment group at baseline and year 1 in the DPP. $^*p < 0.05$ for test of difference between baseline and year 1 glucose value. Only participants who underwent OGTT testing at year 1 are included (i.e., those who did not develop diabetes at 6 months)

Important features of our motivating example include:

- Subjects were required to have 2 samples for confirmation of diabetes usually within 42 weeks so we may be able to assess whether the association between the fasting and 30 minute glucose values change after diagnosis of diabetes so that we may validate the results against these post diabetes values.

- Some subjects were confirmed during a midyear visit when only fasting and not OGTT results are available. For these subjects, there is an additional OGTT collection after the confirmation of diabetes.

- The study showed a significant reduction in the development of diabetes among the treated groups over 10 years of average follow-up (DPP Research Group, 2009) so that the rates of missingness in the collection of OGTT differed across groups. After 6 years of diabetes diagnosis during follow-up, 47% of the Placebo and 36% of the Lifestyle subjects have missing OGTT.

1.3 Notation

We now define the notation used throughout the dissertation by dimension or function under the assumption of a multivariate normal distribution. All vectors and matrices are in boldface and denoted by lower case and upper case letters, respectively. Each subject's data is expressed below with the corresponding form for all s subjects when the index i is dropped.

Indices and scalar quantities

s	number of subjects
t	number of time points
p	number of fixed effects
q	number of random effects
i	index for **subject** with $i = 1, \ldots, s$
j	index for **time** with $j = 1, \ldots, t$
k	index for the explanatory covariates with $k = 1, \ldots, p$
j^*	index for **time** when the event occurred (i.e., when subject i reaches the clinical outcome such that subsequent measurements in the outcome vector is missing (i.e., $y_{i(j^*+1)}, \ldots, y_{it}$) is missing)
δ_i	indicator for a subject with event causing the discontinuation of outcome measurements
n_i	denotes the number of expected outcome measurements per subject i

N denotes the total number of observations in the expected outcome vector $\boldsymbol{y}$ such that $N = \sum_{i=1}^{s} n_i$

Vectors for each subject

$\boldsymbol{y}$ represents the outcome vector of interest with element $\{y_{ij}\}$ and dimension $(N \times 1)$

$\boldsymbol{u}$ represents the vector of time varying auxiliary covariates with element $\{u_{ij}\}$ and dimension $(N \times 1)$

$\widetilde{\boldsymbol{y}}$ represents the concatenated outcome vector $\boldsymbol{y}$ and auxiliary covariate vector $\boldsymbol{u}$ with dimension $(2N \times 1)$

$\boldsymbol{r}$ represents the vector of indicator variables for whether y_{ij} is observed with element $\{r_{ij}\}$ and dimension $(N \times 1)$

$\boldsymbol{y}_{i(o)}$ is the the data vector containing the observed values in $\boldsymbol{y}$ for subject i

$\boldsymbol{y}_{i(m)}$ is the vector containing the missing values in $\boldsymbol{y}$ for subject i

$\boldsymbol{u}_{i(o)}$ is the partitioned vector of the auxiliary covariate values corresponding to the observed outcome vector $\boldsymbol{y}_{i(o)}$ for subject i

$\boldsymbol{u}_{i(m)}$ is the partitioned vector of the auxiliary covariate values corresponding to the missed outcome vector $\boldsymbol{y}_{i(m)}$ for subject i

$\boldsymbol{\theta}$ denotes the vector of parameters that include fixed effect coefficients $\boldsymbol{\beta}$, random effect coefficients, $\boldsymbol{b}_i$ and unique elements of the covariance matrices $\boldsymbol{R}$ and $\boldsymbol{G}$

$\boldsymbol{\Psi}$ is the vector of unknown parameters for the pdf of $\boldsymbol{y}$

Matrices

$\boldsymbol{X}_i$ represents the design matrix for all fixed effects with element $\{x_{ijk}\}$ and dimension $(N \times p)$ with each fixed effect expressed in its appropriate polynomial form

$\boldsymbol{Z}_i$ is the design matrix for all random effects

1.4 Simulation study example

We will now use a small simulation to illustrate the problem that resulted in invalid conclusions if we ignore the missing values. The simulation study consists of the outcome variable $\boldsymbol{y}$, and a correlated auxiliary variable $\boldsymbol{u}$ that is measured at the same time as $\boldsymbol{y}$. In this example, the auxiliary variable also determines the discontinuation so that once

the auxiliary covariate u exceeds some threshold, h, all subsequent outcome measurements are discontinued. In the case of our motivating example, the annually measured fasting and 30 minute glucoses represent $\boldsymbol{u}$ and $\boldsymbol{y}$ respectively and once the fasting glucose becomes 126 or greater, the 30 minute glucose is no longer collected in subsequent visits. To mimic the motivating example, we simulated $\boldsymbol{y}$ and $\boldsymbol{u}$ as data from a multivariate normal distribution defined in equation (1.1).

$$\tilde{\boldsymbol{y}}_i = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim \mathcal{MVN}\left(\boldsymbol{X\beta}, \begin{bmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_{uy} \\ \boldsymbol{\Sigma}_{uy} & \boldsymbol{\Sigma}_u \end{bmatrix}\right) \tag{1.1}$$

The submatrix $\boldsymbol{\Sigma}_y$ reflects the covariance of the outcome vector $\boldsymbol{y}$ which includes the variance at each visit in the diagonal, $var(y_{ij})$, and the covariance between any 2 visits for a given subject, $cov(y_{ij}, y_{ik})$ with $j \neq k$. The submatrix $\boldsymbol{\Sigma}_{yu}$ (and its transpose $\boldsymbol{\Sigma}_{uy}$) reflects the covariance of the outcome of interest and the auxiliary variable at a visit, $cov(y_{ij}, u_{ij})$, or any 2 visits, $cov(y_{ij}, u_{ik})$, for a given subject. Finally, the submatrix $\boldsymbol{\Sigma}_u$ reflects the covariance of the auxiliary variable vector $\boldsymbol{u}$ which includes the variance at each visit in the diagonal, $var(u_{ij})$, and the covariance between any 2 visits for a given subject, $cov(u_{ij}, u_{ik})$ with $j \neq k$.

We generated data from a multivariate normal distribution for 700 subjects each with 4 visits and randomly assigned to 2 treatment groups, A and B. We used compound symmetric covariance structures for $\boldsymbol{y}$ and $\boldsymbol{u}$ and the same covariance for any pair of y_{it} and u_{it}. The covariance parameters for our simulation study are shown in equation (1.2) on the next page. We specified treatment group differences in outcome means (i.e., $\boldsymbol{\mu}_{y(B)} - \boldsymbol{\mu}_{y(A)} = \mathbf{0.5}$) but no difference in auxiliary covariate means (i.e., $\boldsymbol{\mu}_{u(B)} - \boldsymbol{\mu}_{u(A)} = \mathbf{0}$) for all time points.

$$\boldsymbol{\Sigma}_y = \begin{pmatrix} 1 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_{yu} = \begin{pmatrix} -.6 & -.25 & -.25 & -.25 \\ -.25 & -.6 & -.25 & -.25 \\ -.25 & -.25 & -.6 & -.25 \\ -.25 & -.25 & -.25 & -.6 \end{pmatrix},$$
$$\boldsymbol{\Sigma}_u = \begin{pmatrix} 1 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{pmatrix} \tag{1.2}$$

After the simulated data was generated, we checked the $\boldsymbol{u}$ vector of each subject to find the first time when $u_{it} > 90^{th}$ percentile of the of first time point $t = 1$ for Group A, the reference group. All subsequent values in the outcome vector $\boldsymbol{y}$ are then set to missing to simulate the discontinuation of the measurements after reaching the threshold. Using linear mixed models to account for the covariance among measurements from the same subject, we estimated the means at each time point for Group A and Group B. The model included fixed effects for treatment group, time and interaction for group$\times$time . In Figure 1.2 on the following page, we plot the estimated means for 1000 simulations where each column represents a model while each row represents each treatment group. The simulated model (Column 1) reflects the true distribution of the data as we originally specified. The true mean is constant for all 4 time points for both groups. If we ignore the missing data mechanism, we can see in Column 2 that the conclusions will be incorrect due to the underestimation of the mean for all follow-up time points. If we were to naively take the monotone missing data and adjust for the time varying auxiliary variable $\boldsymbol{u}$ in the model, we see in column 3 that this did not improve the biased estimates from Column 2. This example shows that ignoring the missing data will clearly lead to invalid conclusions and we need to account for the missing data mechanism.

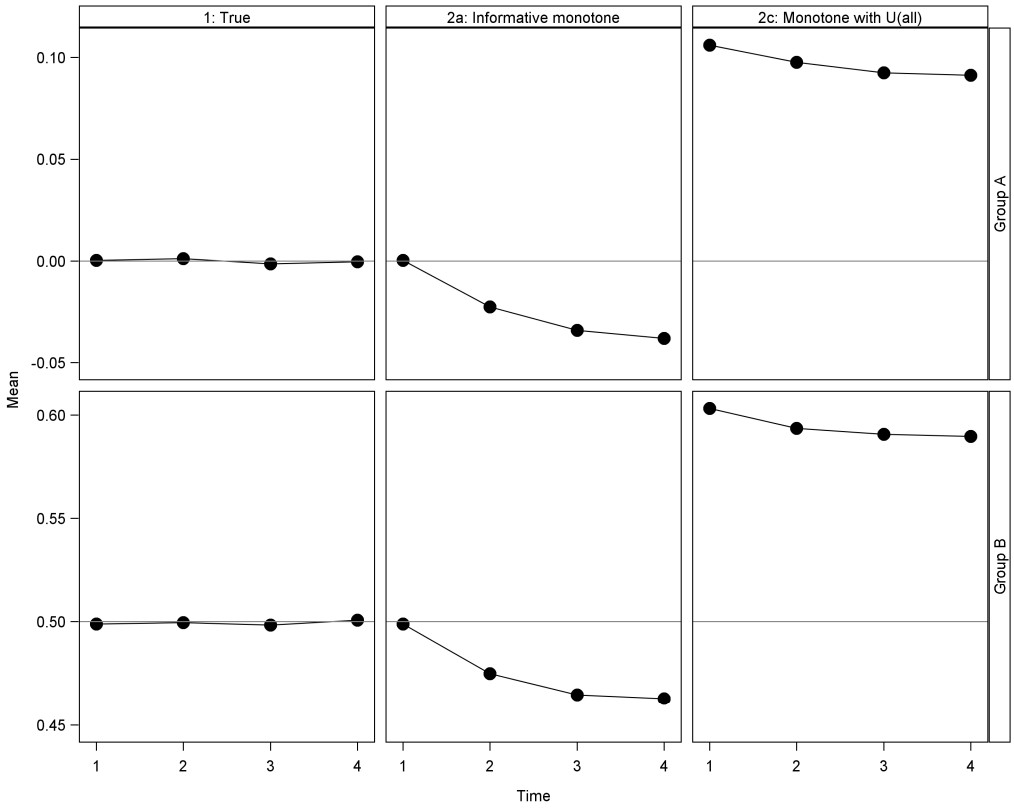


Figure 1.2: Analysis of simulated study using available case analysis. Column 1 represents the true means for both groups. Column 2 involved the analysis of the monotone missing data that assumed missing at random.

1.5 Dissertation Outline

The goal of the dissertation is to propose and study methods to recover the information lost due to the discontinuation of measurements that arose after an absorbing state (i.e., meeting a threshold) occurs. In this dissertation, we review the relevant literature to our problem in Chapter 2 for missing data so that we can characterize the type of missing data at hand and describe possible approaches using the longitudinal data framework for mixed effects models. In Chapter 3, we describe our Imputation-Estimation (IE) based approach to the missing by design problem. We then assessed the performance compared to the other alternatives in Chapter 4. We applied the different estimation methods using

the DPP data from the NIDDK repository in Chapter 5. Finally, in Chapter 6, we summarize our planned approach and describe some possible future directions .

Chapter 2

LITERATURE REVIEW

2.1 MISSING DATA

Missing data in research studies are inevitable especially in longitudinal studies. Unfortunately, missing data can affect the results and conclusions negatively due to bias and inefficiency. In this chapter, we review the definitions, characteristics, and ways to handle missing data from Rubin (1976), Little and Rubin (2002), Schafer and Graham (2002), Daniels and Hogan (2008), Ibrahim and Molenberghs (2009), and Fitzmaurice et al. (2009). The literature on missing data is vast; thus, the review in this section only contains the seminal work and extensions that are directly relevant to the dissertation.

2.1.1 Attributes and Notation for Missingness

We use the following notation to characterize the full data model using the notation from Daniels and Hogan (2008). Let us consider N subjects with measurements at t occasions: the outcome of interest for subject i at time j denoted as y_{ij} ($i = 1, \ldots, N$ and $j = 1, \ldots, t$) and the subject's k^{th} explanatory variable (i.e. age, sex, treatment group, etc.) by x_{ijk} ($k = 1, \ldots, p$). Let ω denote the population of parameters and r_{ij} indicate whether y_{ij} is observed. Examples of targets of inference from the data θ include the effect of explanatory variables on outcomes (i.e., β), or mean differences between groups (i.e.,

$\mu_1 - \mu_2$). The parameter of interest θ indexes the full data response model equation (2.1) which can be specified directly when data are completely observed. Otherwise, the parameter θ is a function of the full data parameter ω indexing the model $p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{x}, \omega)$ so θ is a function of ω.

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \theta) = p(y_1, y_2, y_3, \ldots, y_N \mid \boldsymbol{x}, \theta) \tag{2.1}$$

The full data model can be decomposed into 2 parts: the response model $p(\boldsymbol{y} \mid \theta)$ and the missing data mechanism $p(\boldsymbol{r} \mid \psi)$ as shown in equation (2.2) where $\boldsymbol{y}$ indicates the response vector, $\boldsymbol{x}$ the p-variate explanatory variables, $\boldsymbol{r}$ the vector indicating whether $\boldsymbol{y}$ is observed, and in some cases may include $\boldsymbol{u}$ as the vector of auxiliary variables. This notation will be used later in describing the different missing data mechanisms.

$$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{u}, \omega) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{u}, \theta(\omega))\, p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \psi) \tag{2.2}$$

2.1.2 Types of Missing Data

From sampling survey (Schafer and Graham, 2002), **unit non-response** refers to the situation in which all data from a sampling unit (i.e., individual or school) are missing. **Item non-response** refers to the situation in which information on a particular item or measurement is missing for a subject.

2.1.2.1 Patterns and causes of missing data

Missingness can also be characterized by its pattern in longitudinal studies. The pattern describes which values are missing. Non-response in longitudinal studies may be characterized as monotone, univariate, or wave (Schafer and Graham, 2002). We illustrate the patterns by example in Table 2.1 on the next page where $\boldsymbol{O}$ denotes observed values while $\boldsymbol{M}$ denotes missing values for a subject with outcome vector $\boldsymbol{y}$ with 6 occasions of

measurement.

Table 2.1: Sampling of data patterns

Data Pattern	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	y_{i6}
Complete data	O	O	O	O	O	O
Univariate missing	O	O	M	O	O	O
Wave missing	O	M	O	M	O	M
Monotone missing	O	O	O	M	M	M

A subject may have any one of these sampled profiles. The **complete data** is obviously the goal in a longitudinal study and great care and resources are required in achieving this pattern. **Univariate** missing refers to the situation where only 1 covariate is missing among the set of measurements done on a given visit. In **wave** nonresponse, missing data are scattered in a subject's data vector and may reflect subjects who drop in and out of study visits. The **monotone** pattern has all values missing after the initial missing value which may be attributable to subject drop-out or planned discontinuation of measurements. For monotone missing, once y_{ij} is missing at time j^* for a given subject the remaining $y_{j^*+1}, \ldots, y_t$ measurements are also missing. Subject drop-out is a very important consideration in longitudinal studies and care should be given to identifying the cause of drop out, as we will discuss later. Identifying the pattern for a given study is important because it gives insight on possible causes of missingness.

2.1.2.2 Distribution of missing data and its implications

Rubin (1976) described the distribution (or mechanism) of missingness as it relates to the observed and unobserved components (i.e. $\boldsymbol{x}, \boldsymbol{y}$) of the data. This section describes the different distributions and some methods for identification. The importance lies in whether one can ignore the missing data without compromise to the inferential validity of conclusions. We describe the different types of missing data mechanisms as introduced by Rubin (1976); Little and Rubin (2002) and described by Daniels and Hogan (2008) as $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \psi) = p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{x}, \boldsymbol{u}, \psi)$. The different distributions for missingness are

summarized in Table 2.2 and described below with examples.

Table 2.2: Distribution of missingness indicators

Missing data mechanism	$p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \psi)$
MCAR	$= p(\boldsymbol{r} \mid \psi)$
Covariate MCAR	$= p(\boldsymbol{r} \mid \boldsymbol{x}, \psi)$
MAR	$= p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}, \psi)$
Auxiliary variable MAR	$= p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}, \boldsymbol{u}, \psi)$
MNAR	$\neq p(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}'_{mis}, \boldsymbol{x}, \psi)$

When the missing mechanism is independent of the observed and unobserved outcomes, the distribution of missing is referred to as **missing completely at random (MCAR)**. A special case of this mechanism is MCAR with covariates or covariate-dependent missingness where missingness is still independent of the outcomes but may depend on the explanatory covariates, $\boldsymbol{x}$. Under MCAR, the missing data has the same distribution as the observed data so that the missingness mechanism can be ignored for the analysis of θ but may yield slightly higher variance in the estimation of θ. For example, a randomized clinical trial compares the effect of drug A versus placebo on blood pressure levels after 6 months of intervention. If a randomized subject misses the 6 month evaluation due to family illness, this is an example of missing completely at random since we do not expect family illness to differ between the two treatment groups. For this study, the analyst may ignore the missing data as it should not affect the study conclusions.

When the probability of missing is independent of the unobserved outcomes ($\boldsymbol{y}_{mis}$) and the covariates ($\boldsymbol{x}$), the mechanism is **missing at random (MAR)**. Using a blood pressure example, the systolic blood pressure will not be measured at 6 months if their systolic blood pressure at 3 months is 140 or more mm Hg. If we use the unadjusted mean of SBP at 6 months, the simple mean estimate will be biased. But, if we were to take the conditional mean at 6 months using the 3 month data, the results will be unbiased. A special case of MAR is the **auxiliary missing at random (A-MAR)** where the missing mechanism also depends on an auxiliary variable $\boldsymbol{u}$ instead of the observed outcome $\boldsymbol{y}_{obs}$

(Daniels and Hogan, 2008). Preplanned discontinuation of measurements based on the value of another variable usually falls under A-MAR. For example, a blood pressure study may also collect serum creatinine at the same visit. Serum creatinine may not be the outcome of interest but is associated with blood pressure levels.

When the missing mechanism depends on $\boldsymbol{y}_{mis}$, we have **missing not at random (MNAR)**. For example, in our hypothetical hypertension study, if the 6 month values are missing because the subject was feeling sick due to high blood pressure, this is MNAR. However, it is never possible to reject MAR in favor of MNAR on the basis of the observed data, because we cannot see $\boldsymbol{y}_{mis}$.

2.1.2.3 Ignorability assumption

Rubin (1976) first described conditions where the missing data mechanism can be ignored when making inference for θ. If missing data are MAR and θ and ψ are distinct parameters, then inference based on the likelihood can ignore the missing data mechanism $p\left(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}, \psi\right)$. If in addition, a prior distribution on the parameters factors as $p\left(\theta, \psi\right) = p\left(\theta\right) p\left(\psi\right)$, then the missing data are ignorable for Bayesian inference. That is, Little and Rubin (2002) defines 3 conditions that are required for the ignorable missing data mechanism when making inference regarding the posterior distribution: (1) missing data mechanism must be MAR, (2) ω can be factored as $\omega = (\theta, \psi)$ where θ denotes the parameters from the full data response model and ψ for the missing data mechanism and (3) the decomposed parts of the full data parameter prior distribution are independent, $p\left(\theta, \psi\right) = p\left(\theta\right) p\left(\psi\right)$.

The ignorability assumption allows one to make a valid inference on θ without specifying the missing data mechanism because the posterior inference about θ is proportionally related to a likelihood where the missing data mechanism $p(r \mid y, \psi)$ does not have to be specified (Little and Rubin (2002)):

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_{obs}) \propto p(\boldsymbol{\theta}) p(\boldsymbol{y}_{obs} \mid \boldsymbol{\theta}) \quad \text{or}$$
$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}_{obs}) \propto p(\boldsymbol{\theta}) p(\boldsymbol{y}_{obs} \mid \boldsymbol{x}, \boldsymbol{\theta}) \tag{2.3}$$

The ignorability assumption also implies that the missing responses can be extrapolated from the observed outcome vector $\boldsymbol{y}_{obs}$ as shown in equation (2.4) (Daniels and Hogan, 2008). This important characteristic of the ignorability assumptions allows one to use conditional distribution of the missing outcome $\boldsymbol{y}_{mis}$ given the observed outcome $\boldsymbol{y}_{obs}$ for data imputation under MAR.

$$\begin{aligned} p\left(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{\omega}\right) &= \frac{p\left(\boldsymbol{y}_{mis}, \boldsymbol{y}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}\right)}{p\left(\boldsymbol{y}_{obs}, \boldsymbol{r} \mid \boldsymbol{\omega}\right)} \\ &= \frac{p\left(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{\psi}\right) p\left(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis} \mid \boldsymbol{\theta}\right)}{p\left(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{\psi}\right) \; p\left(\boldsymbol{y}_{obs} \mid \boldsymbol{\theta}\right)} \\ &= \frac{p\left(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis} \mid \boldsymbol{\theta}\right)}{p\left(\boldsymbol{y}_{obs} \mid \boldsymbol{\theta}\right)} \\ &= p\left(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{\theta}\right) \end{aligned} \tag{2.4}$$

2.1.3 Standard Procedures for Missing Data

In many instances, analysis of datasets simply ignore missing data and assumes that the missing data mechanism is MCAR. Unfortunately, this assumption may lead to bias and incorrect inference. Several methods have been proposed in dealing with missing data. The methods fall in one or more of the 5 categories listed below and described in subsequent sections. Applications of the methods are not mutually exclusive and are sometimes used in combination. Ultimately, one should consider the features of the missing data such as the rate at which data are missing (i.e. percent of the data missing), the missing data patterns (monotone versus not), the role of variables in the models (explanatory versus outcome variables), and possible reasons for data to be missing. This evaluation is important as it can help identify the missing data mechanism and the appropriate framework to use. Ultimately, the goal of the procedures outlined below is to

get valid and efficient parameter estimates of interest.

2.1.3.1 Ignore the missingness

Complete case (CC) or available case (AC) analysis falls under this approach. In complete case analysis, only subjects with a complete data vector is used. For example, in a longitudinal study, if any follow-up outcome or explanatory variables are missing, the subject is removed from the analysis. Available case analysis uses all observed data from the subject. For example, in a longitudinal study, if any follow-up outcome or explanatory variables are missing, the subject is still included in the analyses which uses all observed pairs of outcome and explanatory data. In the case of a simple linear regression model, these two approaches are the same since each subject only contributes one observation (i.e., $t = 1$).

In cases when we employ the complete case analysis in randomized trials, the intent to treat principle is violated (Lachin, 2000). The intent to treat principle requires that the analyses of randomized clinical trials include all subjects regardless of protocol adherence or follow-up status to minimize selection bias. Employing a complete case analysis may incur the risk of selection bias. In cases of MCAR, ignoring data are generally valid but may compromise the estimates under MAR due to selection bias resulting in invalid inference.

2.1.3.2 Weight adjustment

Weight adjustment for nonresponse is a common method in survey research and is described in (Little and Rubin, 2002). The idea is that the observed responses are weighted so that the distribution is similar to the population of interest with respect to some auxiliary information. Weights are estimated using the derived probabilities from logistic or probit models and can adjust for the bias arising from the difference in the auxiliary information used. However, the bias arising from the unknown or unspecified covariates are not addressed.

Several schemes exist for how the derived weights are used in the estimation of the parameters. For example, **inverse probability weighting** entails weighing each subject by the inverse of the probability that is observed (Zhao and Lipsitz, 1992; Flanders and Greenland, 1991).

2.1.3.3 Likelihood based approaches

Likelihood based approaches entail partitioning the joint distribution into the full data response model and the missing distribution. Table 2.3 lists three proposed methods for partitioning the likelihood including a special case called a conditional linear model for shared parameter model. All likelihood based methods provide a maximum likelihood estimator for $\boldsymbol{\theta}(\boldsymbol{\omega})$. However, partitioning affects how the distribution of interest $p(\boldsymbol{y} \mid \boldsymbol{x})$ is attained. In selection models (Heckman, 1976), this quantity is directly specified but in pattern mixture models (Rubin, 1977; Little, 1995), $p(\boldsymbol{y} \mid \boldsymbol{x})$ has to be integrated over the different patterns of missingness. Note that the pattern mixture is a reverse factorization of the selection model such that the missingness depends on $\boldsymbol{y}_i$ in the selection model and the data response model is conditional on the missingness $\boldsymbol{r}_i$.

Table 2.3: Decomposition of missingness in likelihood approach

Model Framework	$p(\boldsymbol{y}, \boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{\omega})$
Selection model	$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\omega})p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\omega})$
Pattern mixture model	$p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{x}, \boldsymbol{\omega})p(\boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{\omega})$
Shared parameter model	$\int p(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{\omega})db$
Conditional linear	$\int p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{b}, \boldsymbol{x})p(\boldsymbol{b} \mid \boldsymbol{r}, \boldsymbol{x})p(\boldsymbol{r} \mid \boldsymbol{x})db$

2.1.3.4 Selection Models

Selection models are widely used in the field of econometrics to describe the association between the response rates and sensitive items from a questionnaire such as income level (Heckman, 1976). The selection model partitions the likelihood into the full data distribution, $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\omega})$, and the missing data mechanism, $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\omega})$. The response

may be binary or continuous in these models. In **parametric selection models** (Baker and Laird, 1988), the missing data mechanism are specified by assuming a prior parametric distribution such as Bernoulli or probit for the full data response model and specifying the form of the missing data mechanism as parametric also. In **semi-parametric models**, the missing data are specified using a parametric distribution and the response distribution are derived non-parametrically. If the pattern is monotone for a longitudinal study, the missing mechanism may be modeled as a hazard function. Since the approach uses maximum likelihood estimation (MLE), the estimators are efficient but not robust so care should be taken because of sensitivity to model misspecification.

Daniels and Hogan (2008) described an example where (y_1, y_2) is bivariate normal and y_2 may be missing for subject i. The selection model approach for this example may specify $p(r_i \mid y_{i1}, y_{i2})$ to be distributed as Bernoulli with $\pi(\boldsymbol{\psi})$ as parameter, $Ber\{\pi(\boldsymbol{\psi})\}$, and the missing mechanism is linear in $\boldsymbol{y}$ (i.e., $g(\pi) = \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}$). A special case of the selection model is known as Heckman probit selection model where the $g(\cdot)$ is equal to the inverse normal cdf, Φ^{-1} (Heckman, 1976).

2.1.3.5 Pattern Mixture Models

In a pattern mixture model, the full data distribution is factored so that the subjects are classified into groups according to their missing pattern and the observed distribution is estimated for each group (Rubin, 1977; Little, 1995). For the **pattern mixture model** framework, the full data response is obtained by integrating over the missing indicator distribution, compared to selection model where the full data distribution, $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\omega})$, is directly specified:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\omega}) = \sum_{r \in \mathcal{R}} p(\boldsymbol{y} \mid \boldsymbol{r}, \boldsymbol{x}, \boldsymbol{\omega})\, p(\boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{\omega}) \tag{2.5}$$

For longitudinal pattern mixture models, the groups are classified according to the drop out time and subjects within a group are assumed to have the same distribution of missing values. One of the disadvantages of this model is the non-identifiability of the

parameters arising from the increasing number of mixture patterns as the number of follow-up times get larger.

2.1.3.6 Shared Parameter Models

In **shared parameter models**, the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{r}$ are usually independent conditional on a shared set of latent variables or parameters such as random effects noted by $\boldsymbol{b}$ (Wu and Carroll, 1988). These models have been used in adjusting for selection bias due to non-random drop-out for time to event outcomes as the models allow joint modeling of repeated measures and event times. The model can also be used to model the hazard of an event time as a function of stochastic time varying covariates (Daniels and Hogan, 2008) and thus used primarily in a survival analysis framework.

2.1.3.7 Conditional linear model

The **conditional linear model** is a special class of shared parameter models that may also be expressed as a mixture model (Daniels and Hogan, 2008). The full likelihood further decomposes the joint distribution of the outcome $\boldsymbol{y}$ and missing data mechanism $\boldsymbol{r}$ with the latent variable $\boldsymbol{b}$. Conditional linear models were developed for models with right censoring that cannot be ignored. According to Ibrahim and Molenberghs (2009), right censoring is termed **informative** if the censoring probabilities depend on a subject's underlying rate of change or slopes of the outcome variable. Wu and Bailey (1989) used the conditional linear model to estimate and compare the mean slopes for each group accounting for the censoring times.

2.1.4 Simple Imputation

The imputation approach comes in many different forms, from the simplest (mean imputation) to regression based methods. Sources for imputed values may come from unconditional means, unconditional distributions, conditional means and conditional distributions. We now describe some of these methods with special focus on the regression

based imputation as they will be used in this research. The goal in imputation is to find reasonable values to fill in for the missing data to preserve the distributional shape of the outcome of interest and its relationship to other covariates. The hope is that imputing values will provide valid inference rather than precisely predicting the missing value.

Mean imputation or **imputing unconditional means** entails imputing values with the mean for subject i (i.e., $\bar{y}_{i\cdot} = \frac{\sum_j r_{ij} y_{ij}}{\sum_j r_{ij}}$) or the mean for time j (i.e., $\bar{y}_{\cdot j} = \frac{\sum_i r_{ij} y_{ij}}{\sum_i r_{ij}}$). This naive approach of replacing missing values with the mean underestimates the variance and leads to invalid inference.

Last observation carry forward is used for longitudinal studies where all subsequent missing values are filled in with the last known observation for the subject. This method underestimates the variance like the mean imputation.

Hotdeck imputation Hotdecking is an imputation method where one randomly draws values from observed responses to fill in a nonrespondent's value Schafer and Graham (2002). This method is widely used in sample surveys to preserve the variability in the response. However, it still distorts the correlations with other covariates and measures of association.

2.1.5 Regression Based Imputation

The idea here is to regress the outcome on explanatory covariates for observed cases and predict the missing outcomes. These methods underestimates variability if we pretend the predicted outcome is real.

2.1.6 Expectation Maximization (EM) Algorithm

Dempster et al. (1977) introduced the EM algorithm to compute maximum likelihood estimates that are not available in closed form. The EM algorithm proffers the idea that through an iterative process, one can fill the missing data with the complete data sufficient statistics using the current estimate of unknown parameters and then get an updated estimate of the parameters using the imputed data and observed values. The EM

algorithm is not exclusively used in missing data imputation but also in other models such as unbalanced repeated measures, latent class analysis, and factor analysis.

In general, the EM algorithm involves the following steps.

1. Get initial estimate for parameters from the incomplete and unbalanced data which assumes missing values were MAR
2. E-step: Impute values for the complete data sufficient statistics using estimates based on their conditional expectation.
3. M-step: Obtain MLE of the parameters of the observed and imputed values from the previous step.
4. Iterate between the E and M steps until the parameter estimates converge.

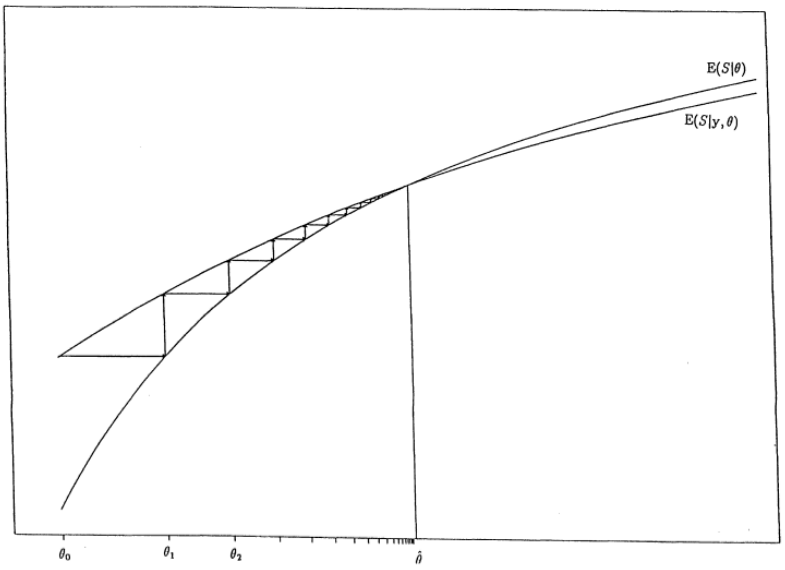


Figure 2.1: Sequence of EM approximations from initial estimate to MLE

The EM algorithm is widely used due to 2 important properties, namely, (1) the EM approximations for a real parameter θ, converges monotonically to the MLE. and (2) the EM sequence increases the likelihood at each step. Navidi (1997) illustrated these properties graphically in Figure 2.1 where $S(x)$ is the sufficient statistics for the complete data. Unfortunately, the EM algorithm is slow converging although further work has improved the convergence rate and with the power of modern computers, has become less of an issue.

2.1.7 Multiple Imputation

Rubin (1987) introduced multiple imputation as an alternative to likelihood based methods such as EM. For each missing value, m simulated values are obtained using an imputation model. Standard analysis methods are then used on the resulting m samples so that each sample will have a separate estimate of the parameters of interest. The mean of the estimated parameters of interest are then used for the overall estimate. Suppose Q and $\sqrt{U}$ represent the parameter of interest and its standard error, the estimated parameter is defined as

$$\bar{Q} = m^{-1} \sum_{j=1}^{m} \widehat{Q}^{(j)} \tag{2.6}$$

and the total variance T consists of the within imputation variance $\bar{U}$ and between imputation variance B:

$$T = \bar{U} + (1 + m^{-1})B \tag{2.7}$$

$$= m^{-1} \sum_{j=1}^{m} \widehat{U}^{(j)} + (1 + m^{-1})(m-1)^{-1} \sum_{j=1}^{m} [\widehat{Q}^{(j)} - \bar{Q}]^2 \tag{2.8}$$

In multiple imputation, the parameters of interest are treated as random instead of fixed so that the posterior distribution is used for drawing the missing data. The Bayesian approach requires one to specify the prior distribution of the unknown parameters but the influence of the assumed priors diminishes with increasing sample size (Schafer and Graham, 2002). The approach also requires specification of 2 models, namely, the imputation and analyst models. The goal of the parametric imputation model is to keep the joint distributions intact by specifying covariates in the model whose relationships should be preserved. Non-parametric methods are also available for the imputation model such as the approximate Bayesian bootstrap (Rubin, 1987). The analyst model defines the explanatory covariates and outcomes to be used to get the estimated parameters of

interest for each of the m samples $(\widehat{Q})^{(j)}$

Many implementations of the multiple imputation procedure are available in SAS, S-plus, R, STATA, SOLAS, and NORM.

2.1.8 Related Developments

Auxiliary information is a covariate that is correlated and sometimes measured at the same time as the response of interest. Collins et al. (2001) conducted simulation studies to contrast the use of an inclusive versus a restrictive strategy for auxiliary variables. Using currently available software, Collins et al. (2001) found that maximum likelihood based methods encourage the use of a restrictive strategy while multiple imputation allows one to be more inclusive which translates to a better possibility in multiple imputation of improved efficiency and reduction in bias.

2.1.8.1 Use of auxiliary information from the same measure

Cook (1997, 2006) used an EM-type of algorithm and multiple imputation methodology to recover the non-ignorable data in a blood pressure study. The missing data arose from cases where a subject was put on antihypertensive medications by their primary physician before getting their final study measurement. To account for the non-ignorable missing mechanism, Cook used a scalar (i.e., the BP value collected by the non-study physician just before the initiation of study meds) which were expected to be correlated with the missing BP. In some cases, the outside BP were not available so they were drawn randomly from the observed outside BP (i.e., multiple imputation). The distribution of the outside BP, the scalar u, was assumed to be a truncated normal in Cook (1997) but the empirical distribution was also examined in Cook (2006). Missing values were imputed with the conditional means after partitioning y into the observed $\boldsymbol{y}_o$ and missing $\boldsymbol{y}_m$ vectors with the assumption that the association between the missing and the auxiliary covariate $\boldsymbol{u}$ is the same as that of the observed and auxiliary covariate. So given the multivariate normal distribution of $\boldsymbol{y}_o, \boldsymbol{y}_m$, and u for each subject,

$$\begin{bmatrix} \mathbf{y}_o \\ \mathbf{y}_m \\ u \end{bmatrix}_{(N+1\times 1)} \sim N\left(\begin{bmatrix} \mathbf{X}_o\beta \\ \mathbf{X}_m\beta \\ \mathbf{X}_u\beta \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_o & \mathbf{\Sigma}_{om} & \mathbf{\Sigma}_{ou} \\ \mathbf{\Sigma}_{mo} & \mathbf{\Sigma}_m & \mathbf{\Sigma}_{mu} \\ \mathbf{\Sigma}_{uo} & \mathbf{\Sigma}_{um} & \sigma_u^2 \end{bmatrix}_{(T+1\times T+1)} \right) \tag{2.9}$$

the conditional expectation for each subject i with a missing data vector $\boldsymbol{y}_m$,is of the form

$$\begin{aligned} E(\mathbf{Y}_m) &= E(\mathbf{y}_m \mid \mathbf{y}_o, u) \\ &= E(\mathbf{y}_m \mid \mathbf{y}_o) + Cov(\mathbf{y}_m, u \mid \mathbf{y}_o)\frac{E(u \mid \mathbf{y}_o)}{V(u \mid \mathbf{y}_o)} \\ &= \mathbf{X}_m\beta \\ &\quad + \mathbf{\Sigma}_m{\mathbf{\Sigma}_o}^{-1}(\mathbf{y}_o - \mathbf{X}_o\beta) \\ &\quad + (\mathbf{\Sigma}_{mu} - \mathbf{\Sigma}_{mo}\mathbf{\Sigma}_o^{-1}\mathbf{\Sigma}_{ou}) \times \frac{u - \mathbf{X}_u\beta - \mathbf{\Sigma}_{uo}{\mathbf{\Sigma}_o}^{-1}(y_o - \mathbf{X}_o\beta)}{\sigma_u^2 - \mathbf{\Sigma}_{uo}{\mathbf{\Sigma}_o}^{-1}\mathbf{\Sigma}_{ou}} \end{aligned} \tag{2.10}$$

with conditional variance of

$$\begin{aligned} Var(\mathbf{Y}_{mi}) &= Var(\mathbf{y}_{mi} \mid \mathbf{y}_{oi}, u_i) \\ &= \mathbf{\Sigma}_m \\ &\quad - \mathbf{\Sigma}_{mo}{\mathbf{\Sigma}_o}^{-1}\mathbf{\Sigma}_{om} \\ &\quad - \frac{(\mathbf{\Sigma}_{mu} - \mathbf{\Sigma}_{mo}\mathbf{\Sigma}_o^{-1}\mathbf{\Sigma}_{ou})(\mathbf{\Sigma}_{um} - \mathbf{\Sigma}_{uo}{\mathbf{\Sigma}_o}^{-1}\mathbf{\Sigma}_{om})}{\sigma_u^2 - \mathbf{\Sigma}_{uo}{\mathbf{\Sigma}_o}^{-1}\mathbf{\Sigma}_{ou}} \end{aligned} \tag{2.11}$$

The value used for the imputation is the conditional mean with an added error term calculated from the Cholesky decomposition of the conditional variance. The Cook model uses the same outcome variable u to get the conditional expectation and does not allow for a different auxiliary measure. The model assumes that the distribution of $\boldsymbol{y}$ and u are the same and that the explanatory variables are the same. Cook showed reasonable performance of this method in the setting of non-ignorable missing data mechanism as long as the model assumptions are correct.

2.1.8.2 Another covariate as auxiliary covariate using multiple imputation

The model of Wang and Hall (2010) allowed for a different auxiliary variable to be used and employed multiple imputation and likelihood based approach to see whether bias can be minimized. Wang and Hall (2010) showed that the likelihood based approach produces consistent and efficient results and slightly more efficient than multiple imputation. However, there is still bias in the estimates when the joint distribution or imputation model is misspecified.

2.1.8.3 Variance adjustments for conditional mean imputation

Schafer and N. (2000) derived a first order approximation to the posterior moments of the parameters as a correction for the undestimation in variance due to the single imputation as an alternative to multiple imputation. When one imputes using conditional mean, $E(y_i \mid \boldsymbol{X}, \boldsymbol{y}_{obs}, \theta) = \mu_i(\theta)$ with associated variance $V(y_i \mid \boldsymbol{X}, \boldsymbol{y}_{obs}, \theta) = \sigma_i^2(\theta)$, there are 2 sources of variability to consider (1) uncertainty in $\boldsymbol{y}_{mis}$ given the imputed means and (2) uncertainty due to estimation of the parameters in the missing data model. Let Q denote the scalar to be estimated from the complete data $(\boldsymbol{X}, \boldsymbol{y})$ using the estimator $\widehat{Q} = (\boldsymbol{X}, \boldsymbol{y})$ with associated estimate of variance for $\widehat{Q}$, $U = U(\boldsymbol{X}, \boldsymbol{y})$. For smooth estimators of the mean where $T_{X_j} = n^{-1} \sum_{i=1}^{n} X_{ij}, T_y = n^{-1} \sum_{i=1}^{n} y_i$ and X_{ij} representing the value of j^{th} covariate for subject i, let

$$\widehat{Q} = g(T_{X_1}, \ldots, T_{X_p}, T_{\boldsymbol{y}})$$

(2.12)

$$V(\widehat{Q} \mid \boldsymbol{X}, \boldsymbol{y}_{obs}) \approx U(X, \boldsymbol{y}_{obs}, \mu(\widehat{\theta})) + C_1 + C_2 \tag{2.13}$$

The corrections C_1 and C_2 correspond to the variability from the 2 sources, the

uncertainty in the imputed means as shown in

$$C_1 = 2\left(\frac{\partial g(\widehat{T})}{\partial T_y}\right)^2 n^{-2} \sum_{i \in mis} \sigma_i^2(\widehat{\theta}) \tag{2.14}$$

and the uncertainty from estimating parameters based on observed as shown in

$$C_2 = \left(\frac{\partial g(\widehat{T})}{\partial T_y}\right)^2 D_\mu(\widehat{\theta})' \Gamma D_\mu(\widehat{\theta}) \tag{2.15}$$

where

$$D_\mu(\widehat{\theta}) = n^{-1} \sum_{i \in mis} \partial \mu_i(\theta)/\partial(\theta)$$

and

$$\Gamma = V(\theta \mid \boldsymbol{X}, \boldsymbol{y}obs)$$

2.1.9 Summary of Missing Data

For our problem outlined in chapter 1, we have item non-response for the missing 30 minute glucose and insulin after diabetes confirmation. For participants who had complete follow-up and diagnosed with diabetes, we expect a monotone missing data pattern in the 30 minute glucose and insulin measurements after the diagnosis. Since the missingness was planned during the design of the study, we can assume an auxiliary missing at random data mechanism (A-MAR) which does not require specification of the missing data mechanism. The ignorability assumption then allows us to use correlated measurements such as the observed outcomes and time varying covariates to extrapolate the $\boldsymbol{y}_m$ using the conditional distribution of $\boldsymbol{y}_{mis}$ given the explanatory covariates and auxiliary variables. In the next chapter, we describe the literature relevant to the analysis

of longitudinal data so that we can have an appropriate model for our data.

2.2 LONGITUDINAL DATA

In our motivating example, we described a study where the subject is measured over time. This is referred to as **longitudinal data** and is one of many types of correlated multivariate data. In contrast, **repeated measures data** entails multiple measurement of the same outcome under different conditions. Other examples of correlated data include data arising from time series, repeated measures, and spatial data (Verbeke and Molenberghs, 2000). In all cases, it is assumed that the sampling units (i.e subject or school) are independent but the correlation arising from measurements within the same unit (i.e subject) must be considered to get efficient estimates for the parameters.

2.2.1 Attributes of Longitudinal Data

In a longitudinal study, subjects are usually measured at scheduled intervals from the start of the study or in the case of a randomized trial, from randomization. **Follow-up time** is defined as the total time a subject has participated in the study and is usually calculated as the difference between the first and last time of measurement expressed in days, months or years. In some cases, subjects who **drop-out** no longer participate in the study due to death or withdrawal from the study.

In longitudinal studies, subjects are followed for a defined duration and seen according to a predefined schedule and number of visits. The **spacing** of the visits refers to the time interval between visits while **balance** describes the distribution of measurements obtained from each subject. The notion of balanced data originated from experimental design where one requires equal number of sampling units assigned to each experimental condition or treatments. In longitudinal data, **balanced** data refers to situations where each sampling unit or subject have the same number of observations. If this condition is not met, then we have an unbalanced design and the analysis framework should account

for the imbalance. For example, MANOVA is not appropriate for unbalanced design since it ignores all subjects who have at least one missing observation.

Time varying covariates are defined as measurements obtained over time for a subject. For example, in the study of blood pressure over time, a time varying covariate may include serum creatinine, or weight. In contrast, **fixed/time invariant covariates** are those characteristics that do not change over the duration of the study such as sex, birth weight, or race/ethnicity.

There are many types of models for longitudinal data which account for the form of the parameters (nonlinear versus linear), type of outcome (continuous, ordinal, binary or categorical), study design and data collected (balance versus unbalanced data), and the goals for inference (mean or variance estimation). For our research, we will assume a multivariate normal distribution and use the linear mixed effects model as a framework. For some continuous outcomes, we may need to consider a transformation for our non-normal outcomes. The next section describes the seminal work of linear mixed models.

2.3 LINEAR MIXED MODELS

The origins of random effect models came from the field of astronomy. Airy (1861) laid the foundation for linear mixed models by describing the errors of observation in astronomy. Subsequently, Fisher developed the theoretical framework within the Analysis of Variance (ANOVA) model using a single random effect that absorbs all unmeasured factors that give rise to different responses from each individual (Fitzmaurice et al., 2009). Under the single random effect ANOVA, only a positive correlation among a subject's measurements are available and only a compound symmetry structure is available. The compound symmetry assumption may not be applicable in some situations such as studies with long term follow-up so that adjacent observations may be more correlated than observations made further apart. Another approach used for longitudinal data is MANOVA (multivariate analysis of variance) which assumes an unstructured covariance structure.

However, this model does not allow unbalanced data and missing data so that the analysis defaults to a complete case analysis as we described in Section 2.1.3.1 on page 17.
While working at the National Institutes of Health, Greenhouse, Halperin and Cornfield introduced the 2 stage approach to the analysis of longitudinal data. These NIH biostatisticians recommended using regression coefficients derived from running least squares models for each subject in the first stage and using the regression coefficients to estimate summary measures (Fitzmaurice et al., 2009). The NIH method did not allow fixed effects in the first stage so that the design matrix x is constrained. However, this method provided the motivations for linear mixed models. The seminal work in mixed effects models was done by Harville (1977) and popularized by Laird and Ware (1982). Their work involved allowing regression coefficients to vary for each subject.
The idea of mixed effects models have been developed for different types of outcomes (continuous, binary, or generalized linear), form of the random or fixed effects (linear versus non-linear), and form of the covariate structure. In subsequent sections of this chapter, we describe the attributes of mixed models and introduce the notation and terminology for linear mixed models as well as the different estimation methods involved. Again, our aim is to describe the seminal work in this vast field and only the extensions related to our proposed research.

2.3.1 Attributes of Mixed Models

In linear mixed models, the association among the repeatedly measured observations are described as the covariance structure or matrix. For longitudinal data, spacing or timing of measurements denotes the time interval between adjacent measurements and will help guide in the selection of the within subject covariance structure. (Wolfinger, 1993, 1996)
For every linear mixed model, one considers whether an explanatory covariate should be **fixed** or **random** and the decision depends on the goal of the model. Covariates are specified as random effects when one believes that the covariate is associated with a probability distribution. Fixed effects are used when one wishes to assess the effect of the

covariate on the outcome. The fixed effects may include any of the explanatory variables of interest, time effects, and any interactions or polynomial form of the time effect, as appropriate.

Fitzmaurice et al. (2009) describes the two classes of linear mixed models, namely, the **marginal** and **linear mixed effects** models. The choice of marginal versus linear mixed effects depends on the goals of the inference. Marginal or population averaged models use only the fixed effects or parameters such that the mean response in terms of the covariates is expressed as $E(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = X_i\boldsymbol{\beta}$. In contrast, linear mixed effects models use both fixed and random effects in the conditional mean of $\boldsymbol{y}_i$, $E(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = X_i\boldsymbol{\beta}^* + Z_i\boldsymbol{b}_i$. In the linear mixed effects model, $\boldsymbol{\beta}^*$ denotes the unit change in outcome on an individual's response compared to $\boldsymbol{\beta}$ which describes the covariate effect on changes in the population mean. So if the goal of a study is to assess the effect of a treatment on changes in an individual's response, one would use the linear mixed effects models. But, if the goal of the study is to assess treatment effects in the overall population, then marginal models are used. Goldberger (1962) showed that substantial gains in efficiency is achieved by using the predicted values (BLUP, best linear unbiased prediction: $\mathbf{X}\beta + \mathbf{Z}\gamma$) instead of the predicted means (BLUE, best linear unbiased estimate: $\mathbf{X}'\beta$) when predicting future observations. In marginal models, three characteristics are specified: (1) the mean response conditional on the covariates, $E(y_{ij} \mid \boldsymbol{X}_{ij}) = \mu_{ij}$; (2) the conditional variance which depend on the mean, and (3) within subject covariance structure to account for the repeated measurements.

2.3.2 Terminology and Notation for Linear Mixed Effects Models

For each subject i at time t, the outcome y_{ij} can be represented by the overall intercept α, subject effect on the intercept b_i, subject effect on time c_i, time effect,overall time effect t_j the design matrix of explanatory covariates x_i with their coefficient β , and errors ϵ_{ij} as in equation (2.16) on the next page. This representation can then be rearranged so that the overall intercept and time effect are absorbed by the subject specific intercept

$(\alpha_i = \alpha + b_i)$ and time effect $(t_i = t_j + c_i)$. To allow for additional subject specific effects, we then generalize the $\alpha_i + t_i$ to $\boldsymbol{z}_{ij}$ for more subject specific random effects as noted by z_{ij} where the subject specific slope has $z_{ij} = 1$.

$$\begin{aligned} y_{ij} &= \alpha + b_i + c_i + t_j + \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{ij} \\ &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + \alpha_i + t_i + \boldsymbol{\varepsilon}_{ij} \\ &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij} \end{aligned} \tag{2.16}$$

This linear mixed model can be expressed in matrix terms for all n subjects, each with n_i measurements for a total of $N = \sum_{i=1}^{s} n_i$ observations with the dimensions noted in equation (2.17).

$$\mathbf{y}_i = \boldsymbol{X}_i\,\boldsymbol{\beta} + \boldsymbol{Z}_i\,\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{2.17}$$

where:

- i serves as the index for the s subjects such that $i = 1, \ldots, s$
- $\boldsymbol{y}_i$ is the $(n_i \times 1)$ vector of outcomes for the i^{th} subject
- $\boldsymbol{X}_i$ is the $(n_i \times p)$ design matrix for the covariates or fixed effects
- $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of population-averaged parameters or fixed effect coefficients with entries $\{\beta_k\},\ k = 1, \ldots, p$
- $\boldsymbol{Z}_i$ is the $(n_i \times q)$ design matrix of random effects. Usually, the random effects are a subset of the $\boldsymbol{X}_i$ matrix to allow for random slopes and intercepts.
- $\boldsymbol{b}_i$ is the $(q \times 1)$ vector of random effects with mean $\mathbf{0}$ and covariance $\boldsymbol{G}$ and each entry denoted as $\{b_{il}\},\ l = 1, \ldots, q$
- $\boldsymbol{\varepsilon}_i$ is the $(n_i \times 1)$ vector of error terms which are independent and have mean $\mathbf{0}$ and covariance $\boldsymbol{R}$
- $cov(b_i, \boldsymbol{\varepsilon}_i) = \mathbf{0}$

In these models, effects of time dependent and independent covariates on longitudinal outcomes can be assessed, the covariance structure of measurements within subject is flexible, and unbalanced or missing data are accommodated. The linear mixed effect

model with normally distributed random variables (i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$ and $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$ is expressed as:

$$\boldsymbol{y}_i \sim \mathcal{N}\left(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}\right) \tag{2.18}$$

Special cases of the LME is the simple random effects model where $\boldsymbol{R} = \sigma^2\boldsymbol{I}_{n_i}$ and the standard general linear model where $\boldsymbol{b}_i = \mathbf{0}$, and $\boldsymbol{G} = \mathbf{0}$. The log likelihood for the linear mixed effects model is expressed as follows without the constant term where the parameters of dimension, $p + q + [p(p+1)/2] + [q(q+1)/2]$ are noted as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{b}', vech'(\boldsymbol{R}), vech'(\boldsymbol{G}))$ and the $vech$ denote the vector of unique elements of the matrix. However, if one chooses to use only the fixed effects (i.e., marginal model), $\boldsymbol{b}$ can be considered as unknown nuisance parameters.

From equation (2.18), the standard log likelihood for the multivariate data shown in equation (2.19) is used in the estimation of parameters in the linear mixed model as described in the next section.

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\left\{\text{N ln } (2\pi) + \sum_{i=1}^{s}\left[\ln \mid \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R} \mid + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\left(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}\right)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right]\right\} \tag{2.19}$$

2.3.2.1 Covariance structure

The covariance structure is characterized by a block diagonal matrix where each element represents the within subject variation among the outcome measurements. The list of common covariance structures is listed below with examples using 4 time points per subject in Table 2.4 on the next page. There are many other types of covariance structures but we principally consider these. The notation will be used in subsequent sections.

- Unstructured - every element in the covariance matrix can differ.
- Compound symmetry - same variance for all time points and the same correlation between any 2 measurements.

- First order autoregressive [AR(1)] - same correlation between any two adjacent time points and correlation decreases by the power of the number of time intervals between the two measures. This covariance structure requires data to be equally spaced.

- Toeplitz - Like the compound symmetry structure, the variance is the same for all time points and similar to the first order autoregressive where correlations are the same for equidistant intervals but do not depend on the number of time intervals between the 2 measures. This covariance structure requires data that are assumed to be equally spaced. In fact, AR(1) is a special case of the Toeplitz.

Table 2.4: Sample covariance structures for 4 time points

Covariance structure	Notation	Example with $t = 4$
Compound symmetry	$CS(\zeta, \rho, t)$	$\begin{pmatrix} \zeta & \rho & \rho & \rho \\ \rho & \zeta & \rho & \rho \\ \rho & \rho & \zeta & \rho \\ \rho & \rho & \rho & \zeta \end{pmatrix}$
First order autoregressive	$AR1(\zeta, \rho, t)$	$\begin{pmatrix} \zeta & \rho & \rho^2 & \rho^3 \\ \rho & \zeta & \rho & \rho^2 \\ \rho^2 & \rho & \zeta & \rho \\ \rho^3 & \rho^2 & \rho & \zeta \end{pmatrix}$
Toeplitz	$TOEP(\{\zeta_1, \zeta_2, \ldots \zeta_t\})$	$\begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 & \zeta_4 \\ \zeta_2 & \zeta_1 & \zeta_2 & \zeta_3 \\ \zeta_3 & \zeta_2 & \zeta_1 & \zeta_2 \\ \zeta_4 & \zeta_3 & \zeta_2 & \zeta_1 \end{pmatrix}$

2.3.3 Estimation of Fixed and Random Effects

Henderson (1984) showed that the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ in $\boldsymbol{\theta}$ is based on the Generalized Least Squares (GLS) using the distribution of $\boldsymbol{y}_i$ in equation (2.18) on the preceding page and can be expressed as equation (2.20) on the next page. Substituting the estimates of

residual variance $\widehat{\boldsymbol{R}}$, and random effect variance $\widehat{\boldsymbol{G}}$ will yield the estimate for the fixed effects coefficients $\boldsymbol{\beta}$. This expression simplifies to the ordinary least squares (OLS) under the special case of the standard general linear model (GLM) where $\boldsymbol{G} = \boldsymbol{0}$.

$$\begin{aligned} \boldsymbol{\beta}_{GLS} &= \left[\sum_{i=1}^{s} \boldsymbol{X}_i'\left(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}\right)^{-} \boldsymbol{X}_i\right]^{-} \left[\sum_{i=1}^{s} \boldsymbol{X}_i'\left(\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}\right)^{-} \boldsymbol{y}_i\right] \\ &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-} \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \end{aligned} \tag{2.20}$$

The solution for the random effect is shown in equation (2.21). Substituting the estimates for the fixed effects coefficients $\boldsymbol{\beta}$, residual variance $\widehat{\boldsymbol{R}}$, and random effect variance $\widehat{\boldsymbol{G}}$ will yield the estimate for the random effect coefficients $\boldsymbol{b}$.

$$\boldsymbol{b}_{GLS} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) \tag{2.21}$$

2.3.4 Estimation of Variance Components

When we do not account for the correlation among measurements from the same subject and use ordinary least squares, estimates of the variance are inflated. In this section, we describe the different methods for estimating the variance parameters from the linear mixed effects model that accounts for the within subject covariance structure. The likelihood based methods for estimation relies on the assumption that $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{b}$ are normally distributed as shown in equation (2.18) on page 33. Both maximum likelihood (MLE) and residual maximum likelihood estimation (REML) methods involves an iterative process that maximizes the log likelihood at each iteration until convergence. The type of covariance structure is first specified so that the parameters can be identified. Table 2.5 on the following page summarizes several common covariance structures and the number of parameters to be estimated.

Table 2.5: Covariance structures

Structure	Description	Number of parameters	(i, j) element
UN	Unstructured	$n(n+1)/2$	$\sigma_i \sigma_j$
CS	Compound Symmetry	2	$\sigma_i \sigma_j \rho$
AR(1)	First order autoregressive	2	$\sigma^2 \rho^{\lvert i-j \rvert}$
TOEP	Toeplitz	No	$\sigma_{\lvert i-j \rvert+1}$

We now discuss the form of these likelihoods that are reparameterized so that only the parameters of interest remains.

2.3.4.1 Maximum likelihood estimates

If we assume that $R = \sigma^2 \boldsymbol{I}$, maximum likelihood estimation involves taking the derivative of the log likelihood shown in equation (2.19) on page 33 with respect to the parameters as shown in equation (2.22).

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sigma^{-2} \sum \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{1}{2} N \sigma^{-2} + \frac{1}{2} \sigma^{-4} \sum (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \\
\frac{\partial l}{\partial D} &= -\frac{1}{2} \sum \boldsymbol{Z}_i \prime \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i - \sigma^{-2} \boldsymbol{Z}_i \prime \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i
\end{aligned} \tag{2.22}$$

The MLE yields unbiased estimates of $\boldsymbol{\beta}$ even in small samples but gives biased variances for finite samples. Several algorithms have been implemented for this iterative process which include the IE algorithm as described in section 2.1.6 on page 21, and the Newton-Raphson method. When the iterative process fails to converge, noniterative methods are also available such as the MIVQUE0 method. The minimum variance

quadratic unbiased estimation (MIVQUE0) is a non-iterative approach based on method of moments estimators for estimating variance components introduced by Goodnight (1978).

2.3.4.2 Restricted maximum likelihood

Laird and Ware (1982) showed that using the generalized least squares residuals in the log likelihood function reduces bias which is known as the **REML**. Although the MLE approach provides consistency, efficiency and asymptotic normality, we prefer to use the REML since it relies on the moment assumptions instead of the stricter normality assumptions Harville (1977). The REML method then entails maximization of the residual log likelihood function $l_R(\widehat{\boldsymbol{\varepsilon}}, \boldsymbol{\theta})$ defined in equation (2.23). Since the $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y}_i - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right)$, the resulting log likelihood for $\widehat{\boldsymbol{\varepsilon}}$ is expressed in matrix notation in equation (2.23). Note that the residual log likelihood equation (2.23) differs from the standard log likelihood by the term $-\frac{1}{2}\, ln \mid \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \mid$.

$$
\begin{aligned}
l_R(\widehat{\boldsymbol{\varepsilon}}, \boldsymbol{\theta}) &= l_R\left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}, \boldsymbol{\theta}\right) \\
&= l(\boldsymbol{y}, \boldsymbol{\theta}) - l\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta}\right) \\
&= -\frac{1}{2}\left\{ N\, ln\,(2\pi) + \left[ln \mid \boldsymbol{V} \mid + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right] \right\} \\
&\quad + \frac{1}{2}\left\{ p\, ln\,(2\pi) + \left[ln \mid \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \mid + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \right] \right\} \\
&= -\frac{1}{2}\Bigg\{ (N-p)\, ln(2\pi) \\
&\quad + \left[ln \mid \boldsymbol{V} \mid + ln \mid \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \mid + \left(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\right)' \boldsymbol{V}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\right) \right] \Bigg\} \\
&\approx -\frac{1}{2}\Bigg\{ (N-p)\, ln(2\pi) \\
&\quad + \left[\ln \mid \boldsymbol{V} \mid + ln \mid \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \mid + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right] \Bigg\}
\end{aligned}
\tag{2.23}
$$

2.3.4.3 General Estimating Equations

Liang and Zeger (1986) introduced **general estimating equations (GEE)** where the

working correlation was designated as a nuisance parameter. The model allows for the within subject correlation without complete specification in the distribution of y_i as long as the first and second moments are specified. The estimating equation for $\boldsymbol{\beta}$ is defined as:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} G_i' V_i^{-1}[\boldsymbol{y}_i - E(\boldsymbol{y}_i \mid \boldsymbol{\beta})] = 0$$

where

$$\begin{aligned} \boldsymbol{y}_i &= n_i \times 1 \text{ vector of outcomes} \\ G_i &= \frac{\partial E(y_i \mid \beta)}{\partial \beta} \\ V_i &= \text{working covariance matrix of } y_i \end{aligned} \tag{2.24}$$

The method uses the sandwich estimator or robust covariance matrix of the $\widehat{\boldsymbol{\beta}}$ which provides asymptotically consistent estimates of the covariance matrix even when the model is misspecified (i.e, the parametric model chosen or the independence of units) as proposed by Huber (1967) and reviewed by Hardin (2003). Huber (1967) showed that if the expected value of the estimating equation has a nonsingular derivative $\mathbf{A}$ at θ_0, then the estimating function T_n of θ_0, $n^{1/2}(T_n - \theta_0)$ is asymptotically normal with mean 0 and covariance expressed as $V_s = A^{-1}BA^{-T}$. In terms of the GEE model, this is expressed as

$$\begin{aligned} &\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \\ &\lim_{N\to\infty} N \left[\sum_{i=1}^{N} G_i' V_i^{-1} G_i\right]^{-1} \left[\sum_{i=1}^{N} G_i' V_i^{-1} \text{cov}(Y_i) G_i V_i^{-1} G_i\right] \left[\sum_{i=1}^{N} G_i' V_i^{-1} G_i\right]^{-1} \end{aligned} \tag{2.25}$$

where $\boldsymbol{V}_i = \phi A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$

It has been shown that there is an increasing inefficiency in the estimate among quasi-likelihood models, incorrectly specified models, and small sample sized-data. Under these conditions, the GEE method does not provide good coverage probabilities.

2.3.5 Related Developments

In this section, we describe the developments we intend to use in this dissertation.

2.3.5.1 Estimation of Covariance

Problems arise when we allow components of the covariance matrix to depend on a covariate because the resulting covariance matrix may not be positive definite (Daniels and Hogan, 2008) in a Bayesian framework. Daniels and Pourahmadi (2002) described a framework for dealing with this problem by using the square root free Cholesky decomposition of the covariance matrix as proposed by Tanabe and Sagae (1992).

2.3.5.2 Use of Kronecker product for covariance structure in multiple outcomes

Galecki (1994) proposed the use of the Kronecker product to estimate the covariance matrix in multivariate repeated measures. Consider a multivariate normal vector $(\boldsymbol{y}, \boldsymbol{U})$ with $var(\boldsymbol{y}) = \boldsymbol{\Sigma}_y$, $var(\boldsymbol{U}) = \boldsymbol{\Sigma}_u$ and $\rho = corr(\boldsymbol{U}) = corr(\boldsymbol{y})$ and $r = corr(y_j, u_j)$. The covariance matrix of $(\boldsymbol{y}, \boldsymbol{U})$ is the Kronecker product of the inter-variable correlation $\boldsymbol{\Sigma}$ and the intra-variable correlation matrix at the same time point $\boldsymbol{R}$:

$$\begin{aligned} \boldsymbol{\Sigma} \otimes \boldsymbol{R} &= \begin{pmatrix} \Sigma_{\boldsymbol{y}} & \Sigma_{\boldsymbol{yu}} \\ \Sigma_{\boldsymbol{uy}} & \Sigma_{\boldsymbol{u}} \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{\boldsymbol{y}} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & \Sigma_{\boldsymbol{yu}} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \Sigma_{\boldsymbol{uy}} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & \Sigma_{\boldsymbol{u}} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \end{pmatrix} \end{aligned} \tag{2.26}$$

Thiebaut et al. (2002) used the Kronecker product as calculated by PROC MIXED in SAS to provide an estimate of the covariance matrix from a bivariate linear mixed model example as described below.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\gamma_i + W_i + \epsilon_i \tag{2.27}$$

Lin et al. (2000) described a model with multiple continuous outcomes and used LMM to derive estimates of the parameters. Roy et al. (2003) extended the model by using

unconditional SD and GEE. These models allowed estimation of both global and individual effects of exposure variables on multiple continuous outcomes by transforming all outcomes to have the same scale.

2.3.6 Summary of Longitudinal Analysis

We intend to use a linear mixed model for our analysis framework as it accounts for the correlation of measurements taken at different time points for a subject. This analysis approach is both frequentist and likelihood based. We also considered the frequentist based GEE as an alternative but the strong assumption of MCAR precluded further consideration.

Chapter 3

IMPUTATION ESTIMATION ALGORITHM USING AUXILIARY COVARIATES

In this chapter, we describe our method for using time varying variables in longitudinal models using a regression based imputation when data are missing by design. We used the approach developed by Cook (2006, 1997) of imputing the missing values with the conditional expectation of the missing outcome given the observed outcome and the auxiliary scalar value described in Section 2.1.8.1 on page 24. In our problem, the auxiliary covariate is not measuring the same entity so the algorithm cannot assume the same distribution for the auxiliary covariate and outcome. We expanded the conditional distribution to include an auxiliary covariate vector instead of a scalar. Assuming a multivariate normal for the joint distribution of the outcome and auxiliary covariate, we derived the conditional distribution of the missing outcome given the observed outcomes and the auxiliary covariate. The linear mixed model provided parameter estimates employed in calculating the conditional expectation and variance.

As we were working on this problem, Wang and Hall (2010) published an approach that uses auxiliary information under a direct likelihood method and multiple imputation to correct for bias when there is informative loss to follow-up (described in Section 2.1.8.2 on page 26) but only considered the problem under a one sample scenario for a longitudinal study. In our research, we use auxiliary covariates to improve prediction for imputed

values and are interested in a clinical trial setting where we would like to assess differences between groups and how possible differential treatment effects in both outcome and auxiliary covariate may affect inference. Under a missing by design scenario, we also consider how different partitions of the auxiliary covariate perform.

3.1 Objectives

In Chapter 1, we described the motivating example from a diabetes prevention study where measurement of the outcome variable is discontinued after a correlated measurement exceeds some threshold. The simulation study example from Chapter 1 will be used in exploring the issues arising from the missing data for the remainder of this proposal. Using simulated data, we found that when we naively ignore the missing data problem and use the observed data in a linear mixed model framework to account for the within subject correlation, the means were underestimated. We will continue to use the notation outlined from Chapter 1 and for simplicity, we will refer to the subject data without the subject indicator i.

Our goal in this research is to assess whether an iterative imputation based approach can improve the estimates for the mean so that conclusions from the treatment group comparisons will be less biased. Specifically, the parameter of interest is the treatment group difference as outlined in equation (3.1).

$$H_0 : \boldsymbol{\mu}_{y(B)} - \boldsymbol{\mu}_{y(A)} = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_{y(B)} - \boldsymbol{\mu}_{y(A)} \neq 0 \tag{3.1}$$

We will also assess the performance in terms of a one sample estimation problem similar to Wang and Hall (2010) under a longitudinal study.

3.2 Missing Data Mechanism

Here, we describe the rationale for the missing data mechanism and implication for prediction of the missing data by example. Let x_1 and x_2 be observed values of fasting

glucose with x_2 happening after the first high value of fasting glucose and x_1 happening up to and including the first high value of fasting glucose. Similarly, let y_1 and y_2 be values of 30-minute glucose with y_2 not observed because it occurs after the first high value of fasting glucose. Then one can write

$$P(y_{(2)}\ missing \mid x_{(1)}, x_{(2)}, y_{(1)}) = P(y_{(2)}\ missing) \mid x_{(1)}) \tag{3.2}$$

which corresponds to a missing at random (MAR) assumption. Daniels and Hogan (2008) call this Auxiliary Missing at Random (A-MAR). In this formulation, which corresponds exactly to the motivating example, missingness in $y_{(2)}$ is not influenced by unknown parameters. As a result, the missingness mechanism is ignorable for likelihood inference. In terms of efficiency, $x_{(1)}$ and $y_{(1)}$ are positively correlated as observed in the actual data. It is beleived that $x_{(2)}$ and $y_{(2)}$ would also be positively correlated. Using $x_{(2)}$ in the model to predict missing $y_{(2)}$ should lead to better prediction for the missing data and better inference for unknown model parameters.

Thus, it seems that we can reasonably assume that the resulting monotone missing pattern falls under Auxiliary Missing at Random (A-MAR) data mechanism from Daniels and Hogan (2008) to invoke the ignorability assumption from Rubin (1976). In our motivating example, the study discontinued measurement of the outcome variable (annual 30 minute glucose) after diabetes was diagnosed (i.e., fasting glucose exceeded the threshold).

The ignorability assumption allows one to ignore the missing data distribution when conducting inference because the missing data mechanism does not depend on the missing outcome. As a consequence of the ignorability assumption, the missing 30 minute glucose can be extrapolated when one conditions on the observed 30 minute glucose and auxiliary covariate, fasting glucose under A-MAR. As stated in Table 2.2 on page 14, the distribution of missingness can be expressed as $p\left(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}, \psi\right) = p\left(\boldsymbol{r} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}, \boldsymbol{u}, \psi\right)$. Since the study collected fasting glucose semi-annually after discontinuation of outcome measurements, we are able to use this "extra" information to extrapolate for visits after diabetes diagnosis. We expect that the dependence between outcome and auxiliary

covariate is stronger for measurements taken at the same time compared to different time points.

3.3 Distribution of Outcome and Auxiliary Variables

Fitzmaurice et al. (2009) described several frameworks for jointly modeling more than one longitudinal measures: multivariate marginal models, conditional models, shared parameter models, and random effects models. We employ both the multivariate marginal model and random effects model for our algorithm. The IE algorithm uses parameters estimated from the joint likelihood of outcome and auxiliary covariates which allows for errors in both the outcome and auxiliary covariate akin to measurement error models (Wu, 2010).

3.3.1 Joint Distribution

We assume that the outcome vector and auxiliary variable is jointly distributed as multivariate normal with parameters defined in equation (3.3).

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix}_{(2n_i \times 1)} \sim \mathcal{MVN}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{y}} = \mathbf{X}_{\mathbf{y}}\boldsymbol{\beta}_{\mathbf{y}} \\ \boldsymbol{\mu}_{\mathbf{u}} = \mathbf{X}_{\mathbf{u}}\boldsymbol{\beta}_{\mathbf{u}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{y}} & \boldsymbol{\Sigma}_{\boldsymbol{yu}} \\ \boldsymbol{\Sigma}_{\boldsymbol{uy}} & \boldsymbol{\Sigma}_{\boldsymbol{u}} \end{bmatrix} \right) \tag{3.3}$$

Note that the design matrix $\boldsymbol{X}$ may differ for the outcome variable $\boldsymbol{y}$ and auxiliary covariate $\boldsymbol{u}$. We allow for different parameters for $\boldsymbol{y}$ and $\boldsymbol{u}$ unlike in Cook (2006) where the correlated covariate was measuring the same mechanism (i.e., blood pressure) and assumed to have the same distribution as the outcome variable.

3.3.2 Conditional Distribution

As a consequence of our ignorable assumption on the mechanism of missingness, we can extrapolate the missing data $\boldsymbol{y}_{i(m)}$ from the observed measurements as we previously noted in equation (2.4) on page 16. Let subject i have monotone missing measurements in $\boldsymbol{y}_i$ starting at time j^*. We can reorder and partition the concatenated vector $\tilde{\boldsymbol{y}}_i$ and the corresponding design matrix for a given subject i into missed and observed. The resulting form for the vector is given in equation (3.4). Two cases for $\boldsymbol{u}_{i(\bullet)}$ can arise in practice. First, $\boldsymbol{u}_{i(\bullet)}$ is only observed for $j \leq j^*$; that is the auxiliary variable is not measured post event j*. In the first case, one cannot gain additional information about $\boldsymbol{y}_{(m)}$ from the auxiliary variable. Second, $\boldsymbol{u}_{i(\bullet)}$ is completely observed like our motivating example where data is missing by design and we have continued follow-up. This is an extension of the approach used in Cook (2006, 1997) from a scalar to a vector of auxiliary covariates $\boldsymbol{u}$. For a subject who had the absorbing state that caused a monotone missing data pattern, we can partition his data vector as follows:

$$\tilde{\boldsymbol{y}}_i = \begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix} = \left[\begin{array}{c} \boldsymbol{y}_{j^*+1} \\ \vdots \\ \boldsymbol{y}_t \\ \hline \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{j^*} \\ \boldsymbol{u}_{i(\bullet)} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y}_{i(m)} \\ \hline \boldsymbol{y}_{i(o)} \\ \boldsymbol{u}_{i(\bullet)} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y}_m \\ \hline \boldsymbol{y}_o \\ \boldsymbol{u}_\bullet \end{array}\right] \tag{3.4}$$

Since we assumed multivariate normal for $\tilde{\boldsymbol{y}}_i$, the conditional distribution has a closed form. The conditional distribution is also multivariate normal. The reconfigured vector will then have the corresponding partitioned parameters defined in equation (3.5) on the following page.

$$\left[\begin{array}{c} \boldsymbol{y}_m \\ \hline \boldsymbol{y}_o \\ \boldsymbol{u}_\bullet \end{array}\right] \sim \mathcal{MVN}\left(\left[\begin{array}{c} \boldsymbol{X}_m\boldsymbol{\beta} \\ \hline \boldsymbol{X}_o\boldsymbol{\beta} \\ \boldsymbol{X}_\bullet\boldsymbol{\beta} \end{array}\right], \left[\begin{array}{c|cc} \boldsymbol{\Sigma}_m & \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{m\bullet} \\ \hline \boldsymbol{\Sigma}_{om} & \boldsymbol{\Sigma}_o & \boldsymbol{\Sigma}_{o\bullet} \\ \boldsymbol{\Sigma}_{\bullet m} & \boldsymbol{\Sigma}_{\bullet o} & \boldsymbol{\Sigma}_\bullet \end{array}\right]\right)$$

(3.5)

3.3.2.1 Conditional Expectation and Variance

We derived the conditional expectation and variance for the partitioned data vector in equation (3.5) expressed in equation (3.6) and equation (3.7)). The conditional expectation of missing $\boldsymbol{y}_{(m)}$ given $\boldsymbol{y}_{(o)}$ and $\boldsymbol{u}_\bullet$ is

$$\begin{aligned} E(\boldsymbol{y}_m \mid \boldsymbol{y}_o, \boldsymbol{u}_\bullet) =& E(\boldsymbol{y}_m \mid \boldsymbol{y}_o) + Cov(\boldsymbol{y}_m, \boldsymbol{u}_\bullet \mid \boldsymbol{y}_o) Var^{-1}(\boldsymbol{u}_\bullet \mid \boldsymbol{y}_o) \left(\boldsymbol{u}_\bullet - E(\boldsymbol{u}_\bullet \mid \boldsymbol{y}_o)\right) \\ =& \boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{\Sigma}_{mo}{\boldsymbol{\Sigma}_o}^{-1}(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}) + \\ & (\boldsymbol{\Sigma}_{m\bullet} - \boldsymbol{\Sigma}_{mo}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{o\bullet}) \left[\boldsymbol{\Sigma}_\bullet - \boldsymbol{\Sigma}_{\bullet o}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{o\bullet}\right]^{-1} \\ & (\boldsymbol{u}_\bullet - \left(\boldsymbol{X}_\bullet\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\bullet o}{\boldsymbol{\Sigma}_o}^{-1}(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta})\right) \end{aligned} \tag{3.6}$$

while the conditional variance is

$$\begin{aligned} Var(\boldsymbol{y}_m \mid \boldsymbol{y}_o, \boldsymbol{u}_\bullet) &= Var(\boldsymbol{y}_m \mid \boldsymbol{y}_o) - Cov(\boldsymbol{y}_m, \boldsymbol{u}_\bullet \mid \boldsymbol{y}_o) Var^{-1}(\boldsymbol{u}_\bullet \mid \boldsymbol{y}_o) Cov(\boldsymbol{u}_\bullet, \boldsymbol{y}_m \mid \boldsymbol{y}_o) \\ &= \boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_{mo}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{om} \\ &\quad - (\boldsymbol{\Sigma}_{m\bullet} - \boldsymbol{\Sigma}_{mo}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{o\bullet})(\boldsymbol{\Sigma}_\bullet - \boldsymbol{\Sigma}_{\bullet o}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{o\bullet})^{-1}(\boldsymbol{\Sigma}_{\bullet m} - \boldsymbol{\Sigma}_{\bullet o}{\boldsymbol{\Sigma}_o}^{-1}\boldsymbol{\Sigma}_{om}). \end{aligned} \tag{3.7}$$

These quantities reflect the average response for a given subject i at time t who had the same evolution of observed outcome and auxiliary covariates. The imputation also

involves using the conditional variance to boost the variability of this marginal response. We also present an equivalent but simpler way of getting the more familiar expressions for the conditional expectation and variance. If we let $\boldsymbol{Z}_m = \boldsymbol{y}_m$ and $\boldsymbol{Z}_o = (\boldsymbol{y}_o, \boldsymbol{u}_\bullet)$ and $\widetilde{\boldsymbol{\Sigma}}$ be the reconfigured covariance matrix, we achieve an alternative formulation where the conditional expectation and variance can then be calculated as equation (3.8) and equation (3.9).

$$E(\mathbf{Z}_m \mid \mathbf{Z}_o) = E(\mathbf{Z}_m) + \widetilde{\boldsymbol{\Sigma}}_{mo}\widetilde{\boldsymbol{\Sigma}}_o^{-1}(\mathbf{Z}_o - E(\mathbf{Z}_o)) \tag{3.8}$$

$$Var(\mathbf{Z}_m \mid \mathbf{Z}_o) = \widetilde{\boldsymbol{\Sigma}}_m - \widetilde{\boldsymbol{\Sigma}}_{mo}\widetilde{\boldsymbol{\Sigma}}_o^{-1}\widetilde{\boldsymbol{\Sigma}}_{om} \tag{3.9}$$

3.4 Parameter Estimation for Longitudinal Models

Fitzmaurice et al. (2009) described approaches to the joint modeling of longitudinal data which include multivariate marginal models, conditional models, shared parameter models, and random effects models. We used two models for our problem, the analysis and imputation model in parallel with the Multiple Imputation approach. As in multiple imputation, one can specify separate models for the analysis and imputation models. For example, demographic variables can be included in the analysis model and not in the imputation model. The multivariate marginal models to get the estimates of the parameters required for the conditional expectation and variance while the linear mixed model analysis approach is used for the analysis model.

3.4.1 Imputation model

We will utilize an iterative approach to get consistent estimates of the conditional mean akin to EM algorithm. Conditional mean imputation was introduced by Buck (1960) as a valid missing data method under MCAR. Little and Rubin (2002) showed that conditional mean imputation is also valid under MAR and describes an approach for incomplete multivariate normal samples which is similar to the method proposed by Cook (2006). In

Buck's method, it was shown that a consistent estimate of the covariance $\mathbf{\Sigma}$ can be constructed by using consistent estimates of the residual variance and covariance of the outcome and regressor. Schafer and N. (2000) showed that imputation by conditional mean from an imputation model results in inferences that are more precise than multiple imputation. The analytic method adjusts the naive variance estimates to reflect uncertainty from two sources: (1) missing outcomes given the imputed means and (2) estimation of the parameters in the missing-data model. For our approach, instead of adjusting the naive estimate, we added the variability in the imputations.

For our algorithm, we followed Cook's approach of adding an error term to the parameter estimates $\boldsymbol{\beta}$ and conditional expectation $E(\boldsymbol{y}_m \mid \boldsymbol{y}_o, \boldsymbol{u}_\bullet)$. This was done to reflect uncertainty as in a Bayesian sense or in accordance with ideas expressed in multiple imputation. At each iteration, the current parameter estimate is modified to reflect its sampling variability. A random normal deviate $g_1 \sim \mathcal{N}(0,1)$ is drawn from each component of the fixed effect parameters $\boldsymbol{\beta}$ and the Cholesky root of the current estimate of the variance of $\widehat{\boldsymbol{\beta}}$ is multiplied. The Cholesky root for $Var(\widehat{\boldsymbol{\beta}})$ is defined as $C'C = Var(\widehat{\boldsymbol{\beta}})$. The parameter value used in the imputation step is $\widehat{\boldsymbol{\beta}} + C'g_1$. The imputed value of $\boldsymbol{y}_{(m)}$ is then determined by equation (3.6) on page 46 and equation (3.6) on page 46 with the modified parameter values plus another normal deviate $g_2 \sim \mathcal{N}(0,1)$ times the Cholesky root of the conditional variance $Cov(\boldsymbol{u}_\bullet, \boldsymbol{y}_m \mid \boldsymbol{y}_o)$.

The goal at the imputation step is to get the best prediction for the missing outcomes using the empirical distributions from the observed outcomes and auxiliary covariates. The **imputation model** will be used to assess the joint distribution of the outcome of interest $\boldsymbol{y}$ and auxiliary variable $\boldsymbol{u}$ equation (3.10).

$$\widetilde{\boldsymbol{y}}_i = \widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{Z}}_i\widetilde{\boldsymbol{b}}_i + \widetilde{\boldsymbol{\varepsilon}}_i, \tag{3.10}$$

where

$$\widetilde{\boldsymbol{X}}_i = \begin{pmatrix} \boldsymbol{X}_i^1 & 0 \\ 0 & \boldsymbol{X}_i{}^2 \end{pmatrix} \qquad \widetilde{\boldsymbol{\beta}}_i = \begin{pmatrix} \boldsymbol{\beta}_i^1 \\ \boldsymbol{\beta}_i^2 \end{pmatrix}$$

$$\widetilde{\boldsymbol{Z}}_i = \begin{pmatrix} \boldsymbol{Z}_i^1 & 0 \\ 0 & \boldsymbol{Z}_i^2 \end{pmatrix} \qquad \widetilde{\boldsymbol{b}}_i = \begin{pmatrix} b_i^1 \\ b_i^2 \end{pmatrix} \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{G}})$$

$$\begin{aligned} var(\widetilde{\boldsymbol{y}}_i) &= \widetilde{\boldsymbol{Z}}_i \widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{Z}}_i' + \widetilde{\boldsymbol{R}} \\ &= \widetilde{\boldsymbol{Z}}_i \widetilde{\boldsymbol{G}} \widetilde{\boldsymbol{Z}}_i' + \boldsymbol{A} \otimes \boldsymbol{\eta} \end{aligned} \qquad \widetilde{\boldsymbol{\varepsilon}}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_i^1 \\ \boldsymbol{\varepsilon}_i^2 \end{pmatrix} \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{R}})$$

This model allows for two classes of models, the multivariate marginal model which ignores the random effects and the linear mixed effects. For example, we can define a bivariate marginal model which requires assumptions about the three marginal associations (among outcome, among auxiliary covariate, and between outcome and auxiliary covariate) which is similar to the model proposed by Galecki (1994) and applied by Thiebaut et al. (2002). Under linear mixed effects model, we can specify random intercepts for subject to model the three marginal associations. The method of generalized least squares described in 2.3.4 on page 35 is used for estimating the covariance parameters. We specify a direct product unstructured covariance matrix for $\boldsymbol{A}$ and $\boldsymbol{\eta}$. The advantage of using this structure is that it allows for different time intervals or spacings between measurements but requires greater number of parameters to estimate. Although we specify an unstructured covariance model for our algorithm, we can further restrict the covariance structure to reduce the number of covariance parameters (s) to estimate using Akaike's information criterion ($AIC = log\ \widehat{L} - \ s$) or the Bayesian Information Criterion ($BIC = log\ \widehat{L} - s\ logN/2$). Models with larger AIC or BIC values are preferred. In some software packages like SAS, AIC and BIC are expressed as negative values so that smaller values are better. Proper modeling of the covariance structure is required to provide valid inferences for fixed effects by improving efficiency. PROC MIXED in SAS allows for restricted covariance structures of bivariate normal distributions using the Kronecker product as described in 2.3.5.2 on page 39. The

Kronecker product covariance assumes that the covariance for a subject $\boldsymbol{R}$ consists of 2 components: the intra-measurement correlation $\boldsymbol{A}$ and inter-measurement $\boldsymbol{\eta}$ correlation as shown in equation (3.11). The Kronecker product of this formulation has been shown to provide symmetric and positive definite matrices.

$$\boldsymbol{A} \otimes \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{A_y} & \boldsymbol{A_{yu}} \\ \boldsymbol{A_{uy}} & \boldsymbol{A_u} \end{pmatrix} \otimes \begin{pmatrix} 1 & \eta_{12} & \eta_{13} & \dots & \eta_{1t} \\ \eta_{21} & 1 & \eta_{23} & \dots & \eta_{2t} \\ \vdots & & \ddots & \dots & \vdots \\ \eta_{t1} & \eta_{t2} & \eta_{t3} & \dots & 1 \end{pmatrix} \tag{3.11}$$

Using the estimated variance components, the estimates for the fixed and random effects are derived using mixed model equations previously described in 2.3.3 on page 34.

3.4.2 Analysis Model

Our primary goal in the analysis model is to estimate the treatment effect efficiently with minimal bias under a single outcome random intercept model. We focus on the marginal class of the linear mixed effects model to provide the framework in the assessment of treatment group differences instead of random effects models which are usually used for the prediction of a future value for a given subject, estimation of variance components or comparison of slopes. We assume that the outcome vector $\boldsymbol{y}_i$ follow a multivariate normal distribution. The framework can easily accommodate both fixed and random effects (possibly a subset of the fixed effects) using the likelihood ratio test for significant random effects described by Verbeke and Molenberghs (2000) to account for the value of the null hypothesis being on the boundary of the parameter space (i.e. variance from the random effect is zero). In our approach, random effects are assumed to be independent of the missing data probability.

For subject i, the linear mixed model is generally expressed as equation (3.12).

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{3.12}$$

where

$$\begin{pmatrix} \boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix} \right)$$

We consider this the **analysis model** where the vector $\boldsymbol{y}_i$ includes only the outcome of interest. The parameters of interest would typically include the primary hypothesis of treatment group difference, $\boldsymbol{\mu}_{y(B)} - \boldsymbol{\mu}_{y(A)}$, and the estimate for the placebo group, $\boldsymbol{\mu}_{(A)}$. In this model, we do not restrict the covariance matrix $\boldsymbol{R}$ to allow the flexibility for unequally spaced measurements. However, we can also assess whether other covariance structures may fit better as described in Wolfinger (1993). For example, AIC may be used to find the best fit for the covariance structure so that we can have valid inference on the fixed effects.

Estimation of the covariance structure, fixed effects and random effects were previously described in Chapter 2. Inferences on the fixed and random effects takes the form of the null hypothesis specified in equation (3.13) as

$$H_0 : \boldsymbol{L} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{pmatrix} = \mathbf{0} \tag{3.13}$$

and corresponding F statistic with $r = \text{rank}(\boldsymbol{L}\widehat{\boldsymbol{C}}\boldsymbol{L})$

$$F = \frac{\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix}' \boldsymbol{L}'(\boldsymbol{L}\widehat{\boldsymbol{C}}\boldsymbol{L}')^{-1}\boldsymbol{L} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix}}{r}$$

(3.14)

where

$$\widehat{C} = \begin{bmatrix} \boldsymbol{X}'\widehat{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\widehat{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\widehat{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\widehat{R}^{-1}\boldsymbol{Z} + \widehat{\boldsymbol{G}}^{-1} \end{bmatrix}^{-}$$

3.5 Imputation-Estimation Algorithm

3.5.1 IE steps

Our proposed algorithm is described in 5 steps below.

Step 0 **Set tolerance limits**

Establish the level of tolerance such that when the difference between the parameter estimates from 2 consecutive steps are within the tolerance limit, convergence is met. The tolerance limit is set separately for the $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ parameters.

Step 1 **Obtain initial estimates of parameters**

Get the initial parameter estimates from the incomplete and unbalanced data. We use PROC MIXED to get the maximum likelihood estimates of the parameters and specify a within subject covariance matrix to be UN@UN to get the Kronecker product for the unstructured case. The model uses all available data and has fixed effects such as time and treatment group. The estimated parameters are then used for the imputation step.

Step 2 **Imputation step**

Add the Cholesky root of the $Var(\widehat{\boldsymbol{\beta}})$ to the estimated coefficients for the fixed effect vector $\boldsymbol{\beta}$ to reflect the sampling variability in the parameter estimates. Use the resulting parameters and the partitioned vector for each subject with missing values to obtain the conditional expectation and variance defined in equation (3.6) on page 46 and equation (3.7) on page 46. Additional variability is added to the conditional expectation using the Cholesky root of the conditional variance equation (3.6) on page 46 and this quantity $E(\widehat{\boldsymbol{y}_m} \mid \boldsymbol{y}_o, \boldsymbol{u}_\bullet)$ is used to impute for the missing $\boldsymbol{y}_m$.

Step 3 **Estimation step**

Using the complete outcome vector $\boldsymbol{y}$ with imputed values from the previous step, obtain updated estimates of the parameters using MLE. After imputing the missing values, we use PROC MIXED to get the updated parameters.

Step 4 **Convergence**

Iterate between steps 2 and 3 until the parameter estimates converge. Convergence is met when difference between the parameter estimates of the fixed effects coefficients $\boldsymbol{\beta}$ and the covariance matrix from 2 iterations are within the specified tolerance level under step 0.

3.5.2 Software

The IE algorithm for auxiliary missing at random (IE4aMAR) developed for this dissertation used SAS 9.2 and is based on the work done by Cook (1997, 2006). As in most biostatistical research, adoption of methods rely on the availability of standard and well documented software. If the proposed research produces good results, we plan to migrate the code to R as we hope to contribute the algorithms in support of this work to the Comprehensive R Archive Network (CRAN) for others to use. We chose R as the future development environment for its depth of computing algorithms and tools for publishing code. The package will include documentation, examples and code.

3.6 Summary and Discussion

We set forth an imputation estimation algorithm to correct for the bias introduced by the discontinuation of outcome measures after a study defined event. The algorithm uses auxiliary covariates measured at the same time as the observed outcome and after the discontinuation. We have informative monotone missing between treatment groups because of the difference in the rates of developing diabetes when a subject's outcome exceeds a threshold. The algorithm allows for flexibility in the spacing of the longitudinal

measures, use of fixed and random effects, and various covariance structure.

Chapter 4

SIMULATION STUDY

4.1 Overview and Goals

Our primary interest is to detect treatment group differences in the outcome variable at the end of the study (time=4). The null and alternate hypotheses for our analysis model is defined in equation (4.1). We also wish to examine the estimated mean for Group A at the last time point ($time = 4$) to assess performance of the IE algorithm in a one sample setting since equal bias in both groups may result in zero bias in the difference between groups.

$$H_0^* : \mu_{y(B)} - \mu_{y(A)} = 0 \text{ versus } H_1^* : \mu_{y(B)} - \mu_{y(A)} \neq 0 \text{ at time} = 4 \tag{4.1}$$

To assess the performance of the Imputation-Estimation algorithm, we conducted simulation studies with varying degrees of association and different hypotheses to compare with the usual available case analysis and multiple imputation. We use bias, MSE and coverage for our performance metrics to identify key features. The simulations were designed to address the questions below.

1. How does the degree of association between the outcome and auxiliary covariate (i.e., $\mathbf{\Sigma}_{yu}$) affect the estimates of the parameters of interest?

2. Is the degree of association among the outcome measures $\mathbf{\Sigma}_y$ more important than

the degree of association among the auxiliary covariates $\mathbf{\Sigma}_u$?

3. Are the correlations among the outcome measures $\mathbf{\Sigma}_y$ more informative than the degree of association between outcome and auxiliary variates $\mathbf{\Sigma}_{yu}$?

4. How does the rate of missing affect the parameter estimates?

Other key questions include: (1) Can the dependence among the observed outcome vector $\boldsymbol{y}_o$ alone correct for the bias?; (2) What form or partition of the auxiliary covariate vector provides the most information for bias correction? (3) How do shifts in the distribution of the auxiliary covariate between treatment group affect the parameters of interest?

4.2 Simulation Methods

Each simulation scenario was repeated 1000 times and contained a sample of 700 subjects equally divided into 2 groups (Group A and Group B), with 4 annual measurements for each subject under a multivariate marginal model. The simulated data were randomly drawn from a multivariate distribution with group specific mean vector and a common covariance matrix $\mathbf{\Sigma}$ using parameters described in the next section. All generated data are assumed to be measured post-baseline so in the models we do not have baseline adjustment. We assumed that the event that causes the monotone missing occurs at a rate of 10 percent per year. We simulate missing by design using the distribution of the auxiliary covariate $\boldsymbol{u}$ for Group A at $time = 1$ as the reference and designating the 90^{th} percentile of the auxiliary variate as cut-point. The absorbing state or event occurs when a subject's auxiliary covariate first exceeds this cut-point so that all outcome values after the event is set to missing. Tolerance level for convergence in the IE algorithm was set at 0.001 for the parameter estimates for $\boldsymbol{\beta}$ and the the covariance estimates $\mathbf{\Sigma}$. For the maximum likelihood estimation under PROC MIXED, we set the maximum number of iterations to 250 to increase the chances for convergence.

We sampled from a multivariate normal since a marginal model implies a linear mixed model with random intercepts and constant residual variance resulting in a compound symmetry structure (West et al., 2007). We demonstrate this by example using a linear mixed model with random slopes (b_1) for time (t_{ij}) and random intercepts (b_0) for subjects in group A as shown in equation (4.2).

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + e_{ij} \tag{4.2}$$

$$\begin{aligned} var(y_{ij}) &= var(b_{0i} + b_{1i}t_j) + var(e_{ij}) \qquad \text{since } \boldsymbol{e}_i \perp \boldsymbol{b}_i \\ &= G_{11} + G_{22}t_{ij}^2 + 2G_{12}t_{ij} + \sigma^2 \end{aligned} \tag{4.3}$$

$$cov(y_{ij}, y_{ik}) = G_{11} + G_{22}t_{ij}t_{ik} + G_{12}(t_{ij} + t_{ik}) \tag{4.4}$$

where

$$\boldsymbol{b}_i = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \right) \text{ and } \boldsymbol{e}_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n_i})$$

If we only consider a random intercept model such that $G_{22} = 0$ and $G_{12} = 0$, the corresponding variance and covariance form the elements of a compound symmetric structure with $var(y_{ij}) = \sigma^2 + G_{11}$ in the diagonal and $cov(y_{ij}, y_{ik}) = G_{11}$ in the off diagonal. In terms of estimation in SAS, marginal models require positive definiteness in $\boldsymbol{\Sigma}$ while linear mixed models require positive definiteness in $\boldsymbol{G}$ and $\boldsymbol{R}$.

Since we are primarily interested in estimating components of the marginal variance $\boldsymbol{\Sigma}$ and not $\boldsymbol{G}$ and $\boldsymbol{R}$, we sampled from the multivariate normal distribution in (3.2) for our simulations. However, we also ran 1000 simulations from the underlying model (4.2) with random intercepts to confirm the equivalence of the 2 approaches (data not shown).

4.2.1 Parameters for the Scenarios

4.2.1.1 Mean parameters

We examined the performance of the different approaches under the null hypothesis of no difference (H_0) in $\boldsymbol{\mu_y}$ and $\boldsymbol{\mu_u}$ between the 2 treatment groups and 3 alternate hypotheses, namely, difference in $\boldsymbol{\mu_y}$ (H_{1y}), difference in $\boldsymbol{\mu_u}$ (H_{1u}), and difference in both $\boldsymbol{\mu_y}$ and $\boldsymbol{\mu_u}$ (H_{1yu}). Because of the difference in $\boldsymbol{\mu_u}$ under H_{1u} and H_{1yu}, the rate of missingness over time varies between the 2 treatment groups. The means examined under the 4 hypotheses are described in Table 4.1.

Table 4.1: Mean vector $\widetilde{\boldsymbol{\mu}} = (\boldsymbol{\mu_y}, \boldsymbol{\mu_u})^T$ under the null and alternate hypotheses considered for the simulation study

	$\widetilde{\boldsymbol{\mu}}$ under H_0	$\widetilde{\boldsymbol{\mu}}$ under H_{1y}	$\widetilde{\boldsymbol{\mu}}$ under H_{1u}	$\widetilde{\boldsymbol{\mu}}$ under H_{1yu}
Group A	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$
Group B	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0.5}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0}\\\mathbf{0.5}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0.5}\\\mathbf{0.5}\end{pmatrix}$
Group B-A	$\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0.5}\\\mathbf{0}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0}\\\mathbf{0.5}\end{pmatrix}$	$\begin{pmatrix}\mathbf{0.5}\\\mathbf{0.5}\end{pmatrix}$

4.2.1.2 Covariance structure

We examined 3 levels of association in each of the covariance 3 submatrices, the association among the outcome of interest ($\boldsymbol{\Sigma_y}$), association among the auxiliary variate ($\boldsymbol{\Sigma_u}$) and between the outcome and auxiliary variate ($\boldsymbol{\Sigma_{yu}}$) (Table 4.2 on the next page). We considered a compound symmetric structure in the submatrices where the variance in the diagonal are unity so that the association between any 2 entries in the off diagonal reflects the correlation. To achieve positive definite covariance matrices, we used 2 levels for the off-diagonal elements of the ($\boldsymbol{\Sigma_{yu}}$). Note that in the simulation studies, we

assumed that the auxiliary and outcome had an inverse relationship. While the correlation between the fasting and 30 minute glucose is positive, it is possible that the outcome and auxiliary variable might be negatively correlated, such as insulin secretion based on the fasting and 30 minute glucose and insulin sensitivity based on the fasting insulin. Thus, the simulation allowed for a negative correlation.. The IE algorithm accommodates both directions (negative and positive) through specifications in SAS. The model specifications allow for unconstrained correlations through the REPEATED statement, whereas using the RANDOM statement restricts the correlations to be positive.

Table 4.2: Covariance parameters for simulations

Levels	$\boldsymbol{\Sigma}_y$ or $\boldsymbol{\Sigma}_u$	$\boldsymbol{\Sigma}_{yu}$ or $\boldsymbol{\Sigma}_{uy}$
$\boldsymbol{L}$: Low	$\begin{pmatrix} 1 & & 0.05 \\ & 1 & \\ 0.05 & & 1 \end{pmatrix}$	$\begin{pmatrix} -.05 & & -.05 \\ & -.05 & \\ -.05 & & -.05 \end{pmatrix}$
$\boldsymbol{M}$: $Moderate$	$\begin{pmatrix} 1 & & .25 \\ & 1 & \\ .25 & & 1 \end{pmatrix}$	$\begin{pmatrix} -.25 & & -.05 \\ & -.25 & \\ -.05 & & -.25 \end{pmatrix}$
$\mathbf{H_{(1)}}$: $High_{(1)}$	$\begin{pmatrix} 1 & & .6 \\ & 1 & \\ .6 & & 1 \end{pmatrix}$	$\begin{pmatrix} -.6 & & -.05 \\ & -.6 & \\ -.05 & & -.6 \end{pmatrix}$
$\mathbf{H_{(2)}}$: $High_{(2)}$	$\begin{pmatrix} 1 & & .6 \\ & 1 & \\ .6 & & 1 \end{pmatrix}$	$\begin{pmatrix} -.6 & & -.25 \\ & -.6 & \\ -.25 & & -.6 \end{pmatrix}$

4.2.2 Simulation Metrics

The goal of this research is to compare methods in analyzing missing data by utilizing time varying auxiliary variables to minimize bias, gain efficiency and minimize the effects

of the missing data mechanism. We will assess the performance of the different methods using the following defined metrics:

1. **Bias** is defined as the difference between the estimate and the true value for the estimated means. Bias will be deemed unacceptable when the absolute size is more than half of the estimate's standard error.

2. **Mean Squared Error (MSE)** is defined as the average distance between the estimate and the true value of the mean and calculated as $MSE = E((\widehat{\mu} - \mu)^2)$. Our goal is to have estimates with MSE closer to zero. Since $MSE = var(\widehat{\mu}) + bias(\widehat{\mu})^2$, the square root of the MSE is equivalent to the standard error for unbiased estimates (i.e., bias=0).

3. **Coverage** represents the percentage of confidence intervals from the simulations that include the true value. Adequate coverage is defined at the 95% level and values below 90% indicate insufficient coverage as this translates to a doubling of the nominal error rate.

4.3 Analysis Model Using Random Effects

A random effects model was employed as the analysis model to estimate the treatment group effects. The 2 treatment groups were assumed to share a common covariance structure which allowed for dependence among annually measured outcome from the same subject. The basic random effects model for the primary analysis is specified for subject i at time t and randomized to group k as

$$y_{ij(k)} = \mu + \alpha_{(k)} + \nu_{t(k)} + \tau_j + b_i + e_{ij} \tag{4.5}$$

where μ is the overall mean, $\alpha_{(k)}$ is the fixed effect for treatment k, $\nu_{j(k)}$ is the fixed effect for treatment at time t, τ_j is the fixed time effect, $b_i \sim \mathcal{N}(0, \sigma_b^2)$ is the random effect for subject i, and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ is the random effect for the measurement for the subject at

time j. The model assumes that the measurement error and random effects are independent. The dependence between any two measurements taken at different times from the same subject was allowed to vary in the variance matrix for subject i and details are described in 3.4.2 on page 50.

4.4 Missing Data Approaches

We considered whether the auxiliary variable can compensate for the missing data and will compare the following methods for dealing with missing data. All missing data methods used the simulated missing data set except for the True method which used the complete simulated data without missingness after a subject reached the absorbing state. The missing data methods considered below used the basic random effects specified in equation (4.5) on the previous page with possible imputations for the outcome vector or adjustments for time varying auxiliary covariate.

1: True This is the simulated data without the missing values and is considered the "gold standard" for comparison among the missing data methods. The basic random effects model was applied.

2a: AC without U

Under available case without auxiliary information, the missing data is treated as missing at random so the basic random effects model above was used without any modifications.

2b: AC with U(dm)

Under available case with auxiliary information, the missing data is treated as missing at random but recognizes that the missingness depends on the auxiliary covariate. This naive approach does not impute for the missing outcomes but adds a fixed effect for the auxiliary covariate. For those with the event, the auxiliary covariate takes the value at the time of the event (i.e, dm) for all time points but uses the time varying auxiliary

values for those without the event.

2c: AC with U

This approach is similar to *2b: AC with U(dm)* in using just the observed data but adds a fixed effect term to the basic random effects model for the time varying auxiliary covariate.

3a: MI without U

The missing values under this method was imputed used a regression based multiple imputation algorithm. The imputation model was specified for the monotone missing pattern with effects for time, subject, and treatment group and an interaction term for time*treatment group. From this model, we can assess whether the association among the outcome measures is sufficient. We used PROC MI and MIANALYZE for this approach with 5 imputations specified. The basic random effects model was applied to the 5 imputation sets to estimate the parameters of interest.

3b: MI with U(all)

In this approach, we extended approach 3a to include the the time-varying auxiliary covariate in the imputation model. The auxiliary information was used in the imputation model to preserve the dependence between auxiliary and outcome variables but was not included in the analysis model since we are only interested in treatment effects.

4a: IE with U(o)

We used the IE algorithm with the observed partition $\boldsymbol{u}_{(o)}$ which corresponds to the measurements taken when the outcomes were observed $\boldsymbol{y}_{(o)}$ to impute for the missing outcomes. The imputed dataset is obtained using the steps specified in section 3.5.1 on page 52. The basic random effects model was then applied to the data with imputed outcomes.

4b: IE with U(dm+m)

This method is the same as *4a: IE with U(o)* except that we used the

partition of the auxiliary variable vector $\boldsymbol{u}$ corresponding to the outcome $\boldsymbol{y}$ at the time of index event and missing measurements (i.e., $u_{j^*}, u_{j^*+1}, \ldots, u_t$).

4c: IE with U(all)

This method is the same as *4a: IE with U(o)* except for the use of the complete auxiliary covariate vector $\boldsymbol{u}$.

5: EBLUP This method imputes the missing values with the empirical best linear unbiased prediction from a linear mixed effects model with random effects for intercept and fixed effects for time, treatment group, auxiliary covariate and the interaction of time and treatment group.

4.5 Simulation Results

4.5.1 Missing Data

The simulated data resulted in monotone missing pattern in the outcome variable $\boldsymbol{y}$ after a subject's auxiliary covariate $\boldsymbol{u}$ first exceeds the cutpoint. The rate of missingness in the outcome variable $\boldsymbol{y}$ depends on the distribution of the auxiliary variable $\boldsymbol{u}$. Table 4.3 on the next page summarizes the rate of missing according to the degree of association among the auxiliary covariate $\boldsymbol{\Sigma}_u$ (defined in Table 4.2 on page 59) and the vector of mean differences in the auxiliary covariate between groups A and B ($\boldsymbol{\mu}_{u(B)} - \boldsymbol{\mu}_{u(A)}$). The rate of missing was zero at year 1 and increased over 4 years in both groups. Treatment group A is the reference group so that the rate of missing did not change between the 2 mean vectors. Because of the shift in the distribution of the auxiliary covariate $\boldsymbol{u}$, treatment group B experienced the greatest rates of missingness (51%) at Year 4 when $\boldsymbol{\mu}_{u(B)} - \boldsymbol{\mu}_{u(A)} = \mathbf{0.5}$ and low $\boldsymbol{\Sigma}_u$. The missing rates for Group B is twice that of Group A under $\boldsymbol{\mu}_{u(B)} - \boldsymbol{\mu}_{u(A)} = \mathbf{0.5}$ resulting in different rates of missing between the 2 groups.

Table 4.3: Rate of missingness in simulated outcome variables by treatment group

		Group A				Group B			
$\boldsymbol{\mu}_{u(B)} - \boldsymbol{\mu}_{u(A)}$	$\boldsymbol{\Sigma}_u$	Year: 1	2	3	4	Year: 1	2	3	4
0	Low	0%	10%	19%	27%	0%	10%	19%	27%
	Moderate	0%	10%	18%	25%	0%	10%	18%	25%
	High	0%	10%	16%	20%	0%	10%	16%	20%
0.5	Low	0%	10%	19%	27%	0%	22%	38%	51%
	Moderate	0%	10%	18%	25%	0%	22%	37%	47%
	High	0%	10%	16%	21%	0%	22%	32%	39%

4.5.2 IE Algorithm Performance

The IE algorithm is known to be slow converging but the IE4aMAR routine we developed seemed to perform well. We maximized the likelihood under a linear mixed model using PROC MIXED from SAS. We sometimes encountered infinite likelihood problems which we solved by drawing another random sample. Under the expectation step, we had some problems with the non-positive definite covariance matrices and when this occurred, we drew another sample. In Table A.1 on page 101, we summarize the mean number of iterations for IE convergence. The mean number of EM iterations ranged from 2 (under Low $\boldsymbol{\Sigma}_{yu}$) to 9 (under High $\boldsymbol{\Sigma}_{yu}$).

4.5.3 General Findings

The summary data from the simulation studies are found in Table A.2 on page 103. The same data in Figures 4.2 on page 67 to 4.5 on page 71, will be the focus of the summaries below. The problem of coverage from the IE algorithm may due to the underestimation of the variance components from singly imputed values. This problem occurs because the algorithm imputes daa and then treats all the data as if it were missing. The estimated standard errors are therefor too small. If we plot the estimated mean and confidence interval for the 1000 samples in Figure 4.1 on the next page under H_0 with covariance scenario of 07 MMM, we see that the variance was under estimated compared to the multiple imputation approach which tended to have greater variance than the true data.

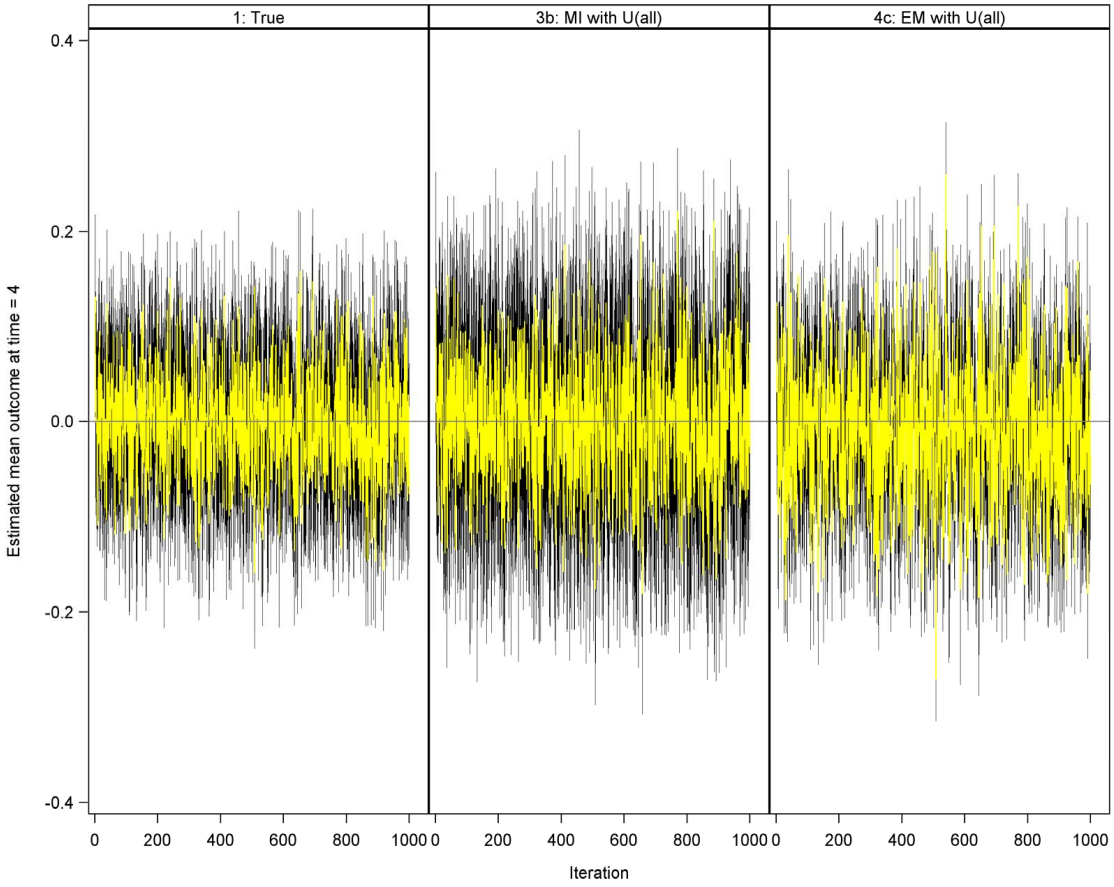


Figure 4.1: Estimated mean $\widehat{\mu}_{y(A)}$ and 95% CI at time=4 for 1000 samples. The yellow lines indicate the estimated mean while the black and blue indicate the 95% confidence band.

Comparison of Estimation Methods Under One Sample Setting

Figures 4.2 on page 67 and 4.3 on page 68 summarize the performance for the 9 estimation methods considered under the null and 3 alternate hypotheses (defined in Table 4.1 on page 58) and 15 combinations of the covariance structures (defined in Table 4.2 on page 59). We assess performance by comparing the bias, MSE and coverage from the 8 methods to the true data denoted by the circle (*1: True*). Among the estimation methods considered, multiple imputation using the auxiliary covariate (*3b: MI with U(all)*) consistently had the best performance (no bias, low MSE, and 95% coverage). However, multiple imputation without accounting for the auxiliary covariate (*3a: MI without U*) resulted in gross deviation from the true value resulting in large negative bias,

large MSE, and very low coverage. Under available case approaches, the deviations from the true value increases as the degree of association between the outcome and auxiliary covariate increases (i.e., $\boldsymbol{\Sigma}_{yu}$). The available case analysis that did not adjust for the auxiliary covariate (*2a: Informative monotone*) performed well under low $\boldsymbol{\Sigma}_{yu}$ since the missingness mechanism is at random. When the outcome measures are highly correlated (i.e., high $\boldsymbol{\Sigma}_{y}$), using the *2a: Informative monotone* data alone results in negative bias and poorer performance. There was no appreciable difference in bias and MSE between estimation under H_0 compared to H_{1y} except for slight differences in coverage in higher degrees of association. Estimation using the available cases adjusted for the auxiliary covariate (*2c: Monotone with U(all)*) tend to overestimate the true value leading to a positive bias, large MSE and poor coverage under H_{1u} and H_{1yu} when association between outcome and auxiliary covariate are moderate and high. Use of the IE algorithm resulted in positive bias for low $\boldsymbol{\Sigma}_{yu}$. The IE algorithm resulted in MSE close to that produced by *3b: MI with U(all)*, but the coverage is acceptable only when $\boldsymbol{y}$ and $\boldsymbol{u}$ are highly correlated (i.e., high $\boldsymbol{\Sigma}_{yu}$).

One of the key questions we posed under a missing by design mechanism was, does the use of missing partition of the the auxiliary covariate ($\boldsymbol{u}_{(m)}$) improve the estimates? When we compare methods 4a to 4c under all hypotheses examined, it seems that there is significant improvement when the association between the outcome and the auxiliary covariate is strong (i.e., high $\boldsymbol{\Sigma}_{yu}$) especially under low $\boldsymbol{\Sigma}_{u}$ and high $\boldsymbol{\Sigma}_{y}$.

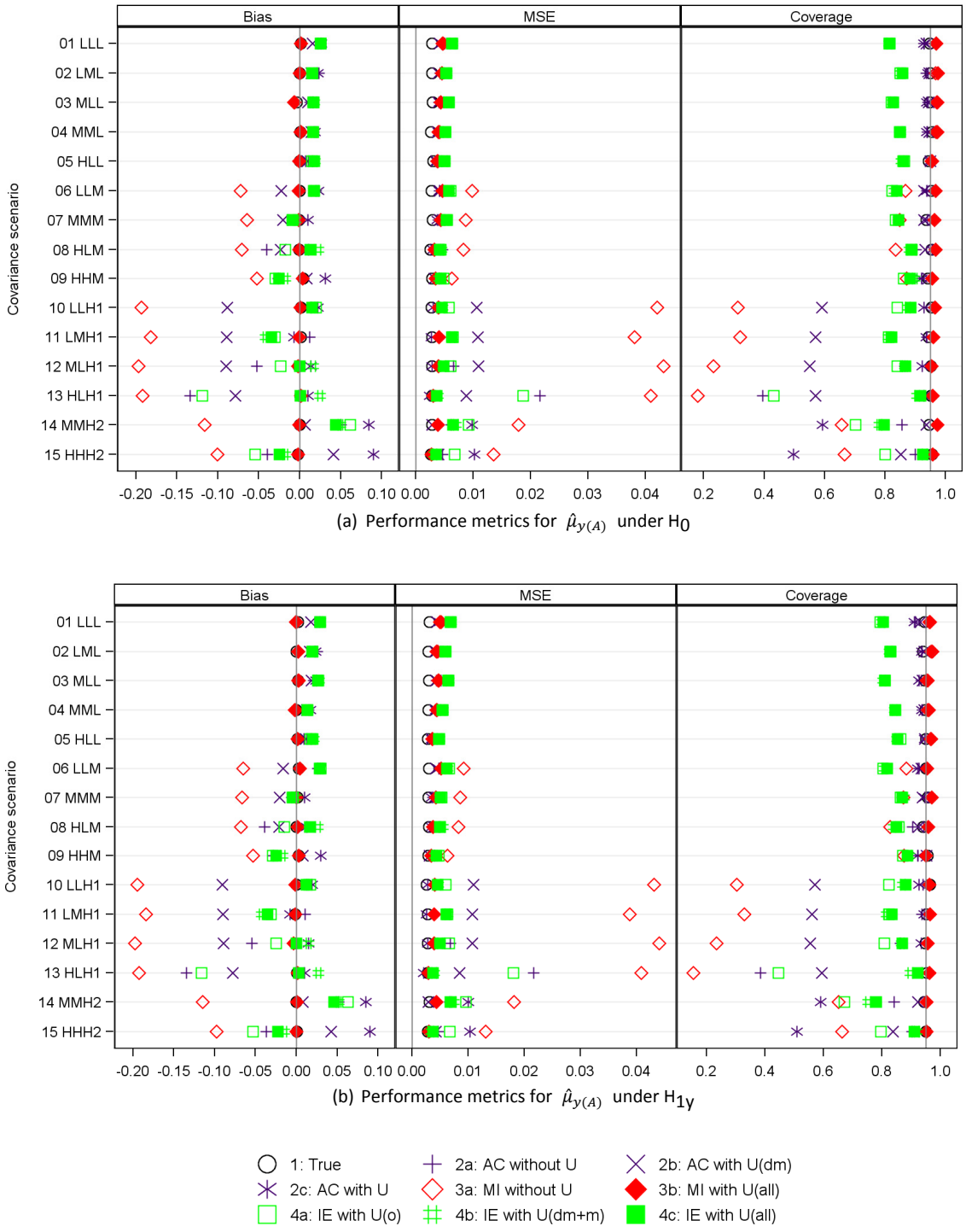


Figure 4.2: Performance for $\widehat{\mu}_{y(A)}$ under H_0 and H_{1y} at time=4

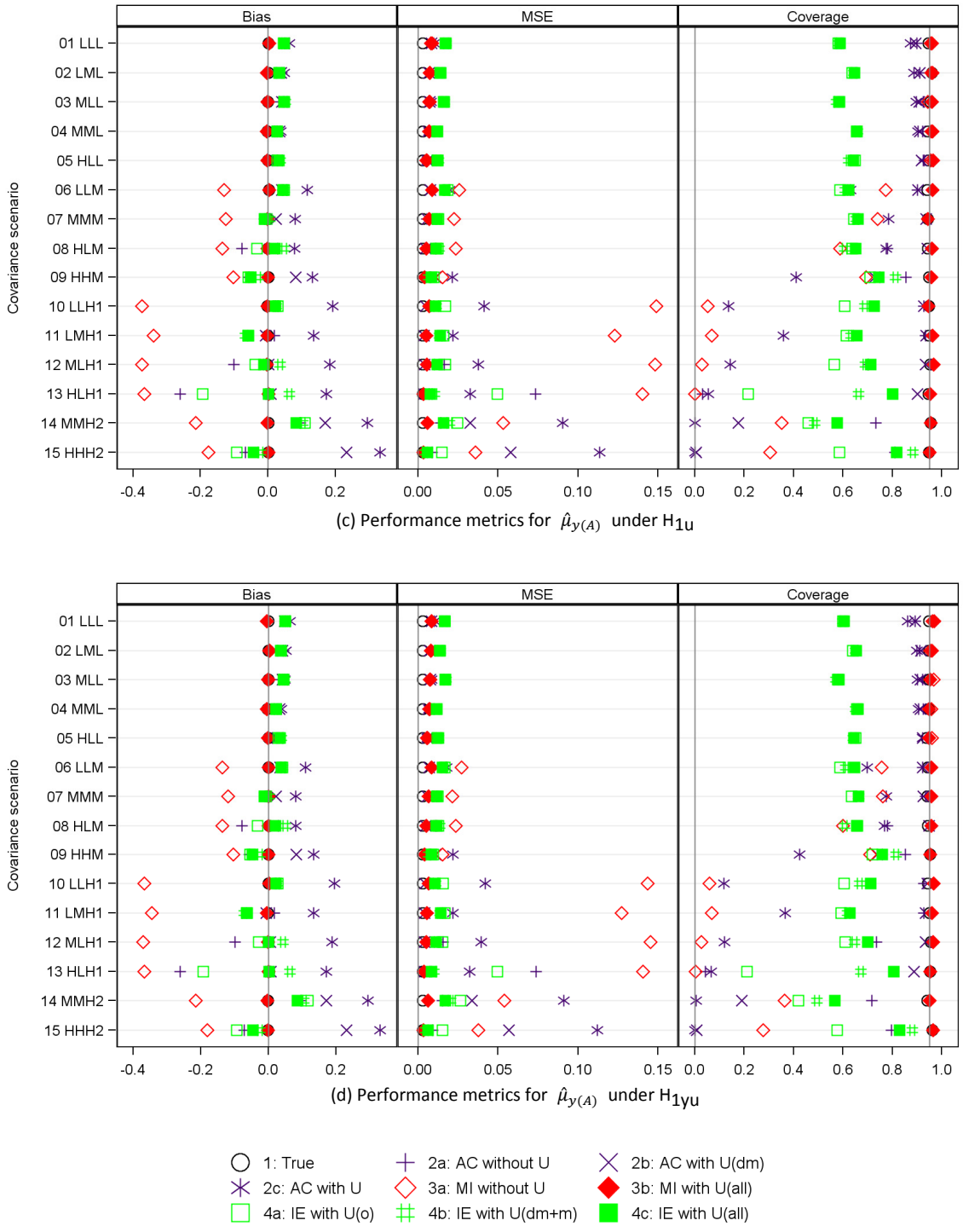


Figure 4.3: Performance for $\widehat{\mu}_{y(A)}$ under H_{1u} and H_{1yu} at time=4

Comparison of Estimation Methods Under Two Sample Setting

Our primary interest is the treatment comparison of the mean in the outcome variable at the last time point, $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$. The bias from the 9 methods does not follow the same pattern as the one seen for $\widehat{\mu}_{y(A)}$ where greater bias was observed for increasing association between outcome and auxiliary covariate. There are negligible differences in performance among all the methods under the H_0 and H_{1y} showing that differences in the outcome means between the 2 groups do not affect the performance. However, under H_{1u} and H_{1yu} the differences in the distribution in auxiliary covariate between the 2 groups result in bias for $\boldsymbol{y}$ due to different rates of missing. Under H_0 and H_{1y} where the distribution of $\boldsymbol{u}$ do not differ between the 2 treatment groups, the available case analysis with adjustment for $\boldsymbol{u}$ has less MSE than multiple imputation. In contrast, the MSE is inflated when the distribution in $\boldsymbol{u}$ differs due to the bias. Under H_{1u} and H_{1yu}, we can see that the MSE is nearly twofold which reflects the different rates of missing between the 2 groups. When the association of $\boldsymbol{y}$ and $\boldsymbol{u}$ is strong (Scenario 1), great improvement in using the $\boldsymbol{u}_{(m)}$ when data are missing by design. Bias is not a problem for $\boldsymbol{\mu}_{(B)} - \boldsymbol{\mu}_{(A)}$ when there is no difference in the distribution of $\boldsymbol{u}$ and the MSE is mainly driven by the resulting variance.

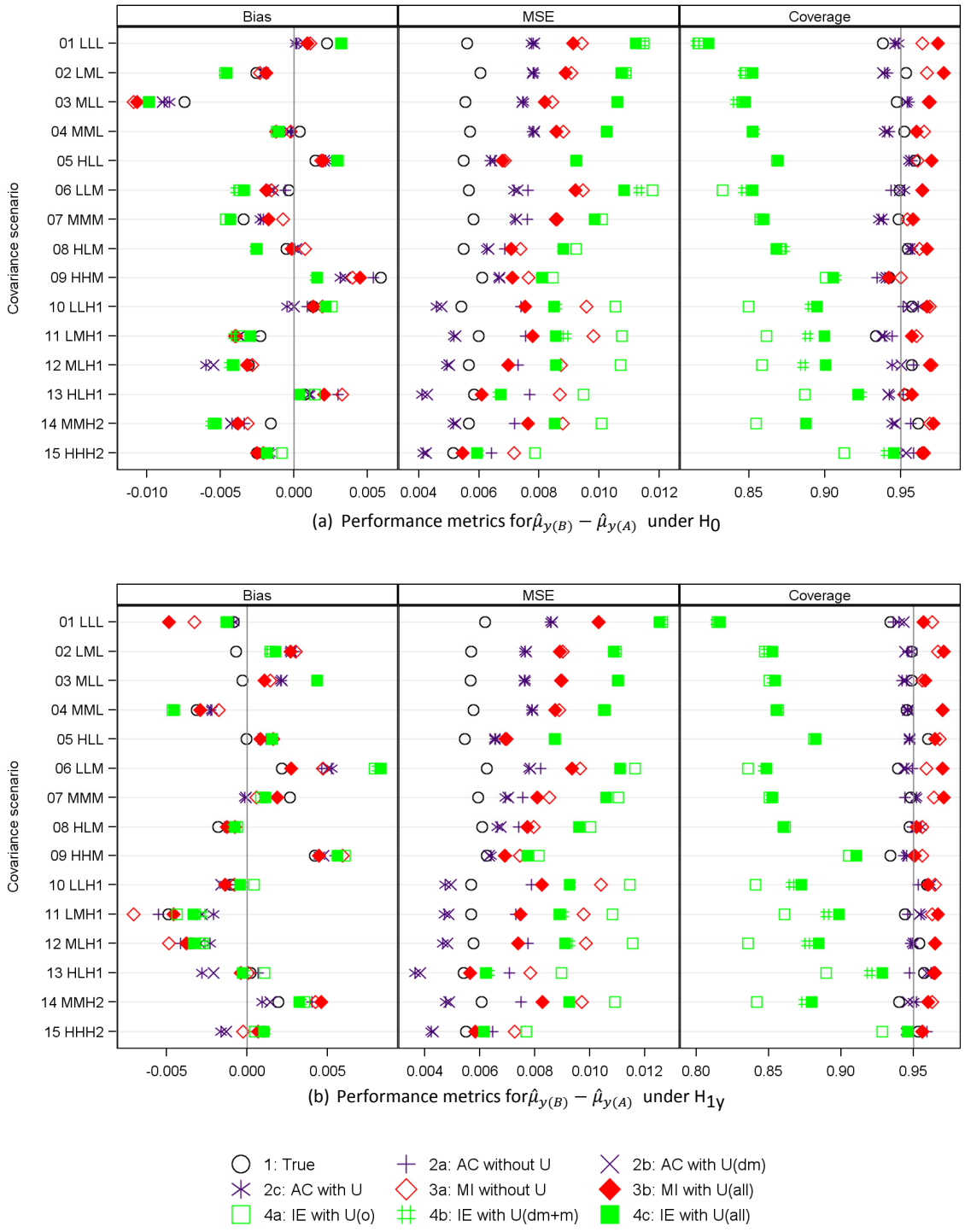


Figure 4.4: Performance for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ under H_0 and H_{1y} at time = 4

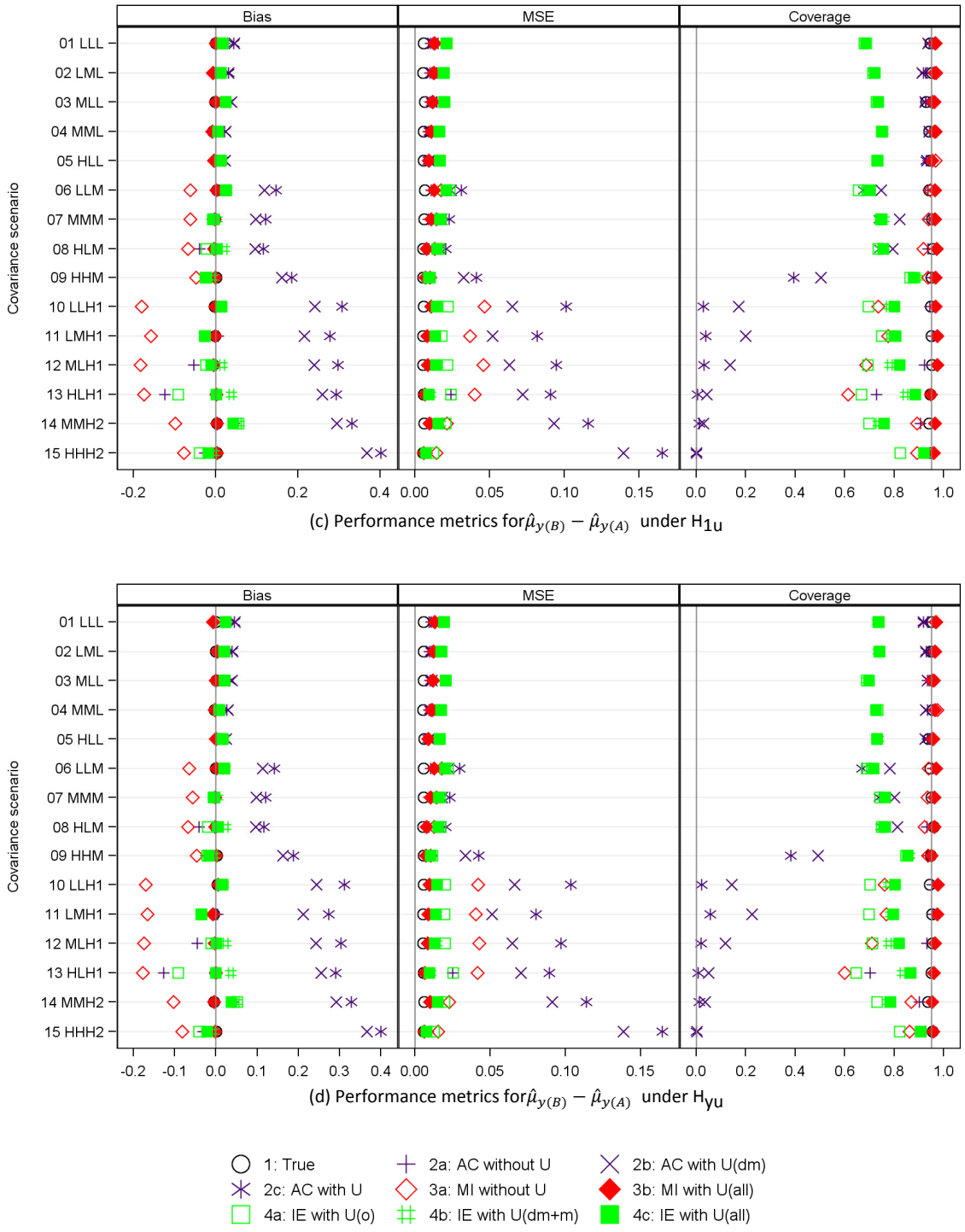


Figure 4.5: Performance for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ under H_{1u} and H_{1yu} at time = 4

4.5.4 Specific Questions

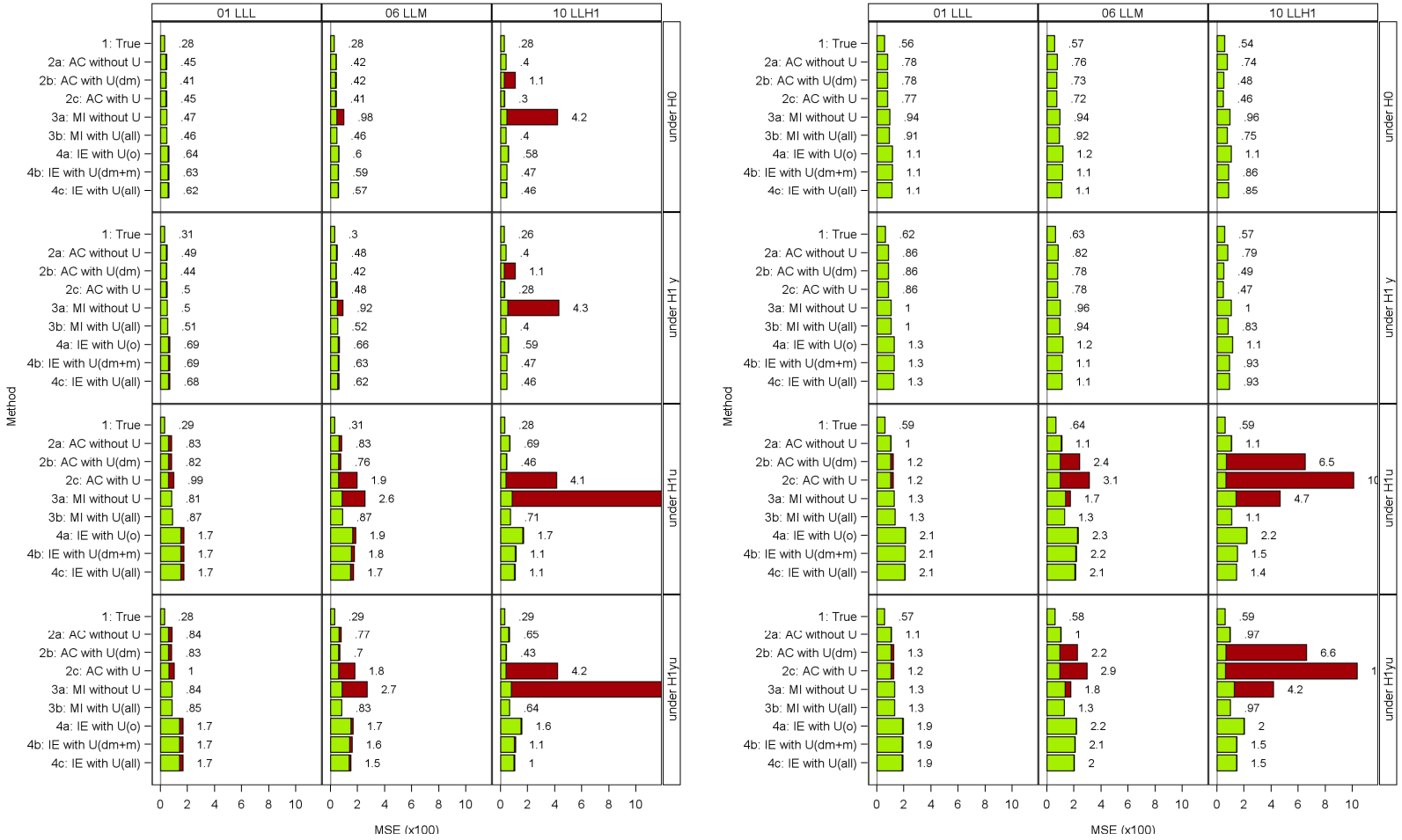
The simulation studies were designed to answer the specific questions outlined below by comparing the MSE of selected scenarios. In the following summaries, we subdivided the MSE to denote the proportion due to the bias versus the variance of the estimate for the parameters $\mu_{y(A)}$ and $\mu_{y(B)} - \mu_{y(A)}$ at time = 4.

Question 1. How does the degree of association between the outcome and auxiliary covariate (i.e., $\boldsymbol{\Sigma}_{yu}$) affect the estimates of the parameters of interest?
In Figure 4.6 on page 74, we contrast 3 degrees of association of the covariance between the outcome and auxiliary covariate $\boldsymbol{\Sigma_{yu}}$, under low $\boldsymbol{\Sigma}_y$ and low $\boldsymbol{\Sigma}_u$. When $\boldsymbol{\Sigma_{yu}}$ is low, bias is not a problem in all of the methods considered since the mechanism is missing at random. The MSE using the available case analysis under scenario 01 LLL is similar to the multiple imputation and nearly double in the EM based methods. When there is no difference in the distribution of the auxiliary covariate between the 2 groups (under H_0 and H_{1y}), we detected bias in $\widehat{\mu}_{y(A)}$ but not in $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ when using the multiple imputation without accounting for the auxiliary covariate. When there is a difference in the distribution of the auxiliary covariate, not accounting for it in the multiple imputation approach results in extreme bias for stronger association in $\boldsymbol{\Sigma}_{yu}$.

Question 2. Is the degree of association among $\boldsymbol{y}$ more important than the degree of association among $\boldsymbol{u}$?
For question 2, we contrast covariance scenarios moderate versus low association in $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_u$ in the presence of high $\boldsymbol{\Sigma}_{yu}$ (Figure 4.7 on page 75). The MSE for multiple imputation approach using the auxiliary covariate did not differ between the covariance scenarios 11 and 12 but improvement was seen when the association among the outcome measures (covariance scenario 13). The MSE from the IE algorithm approaches were also similar between scenarios 11 and 12 with slight improvement with high $\boldsymbol{\Sigma}_y$. There was a marked difference among the 3 EM approaches under scenario 13 where not accounting for

the $\boldsymbol{u}_{(m)}$ vector resulted in substantial bias and higher MSE when the parameter is $\mu_{y(A)}$.

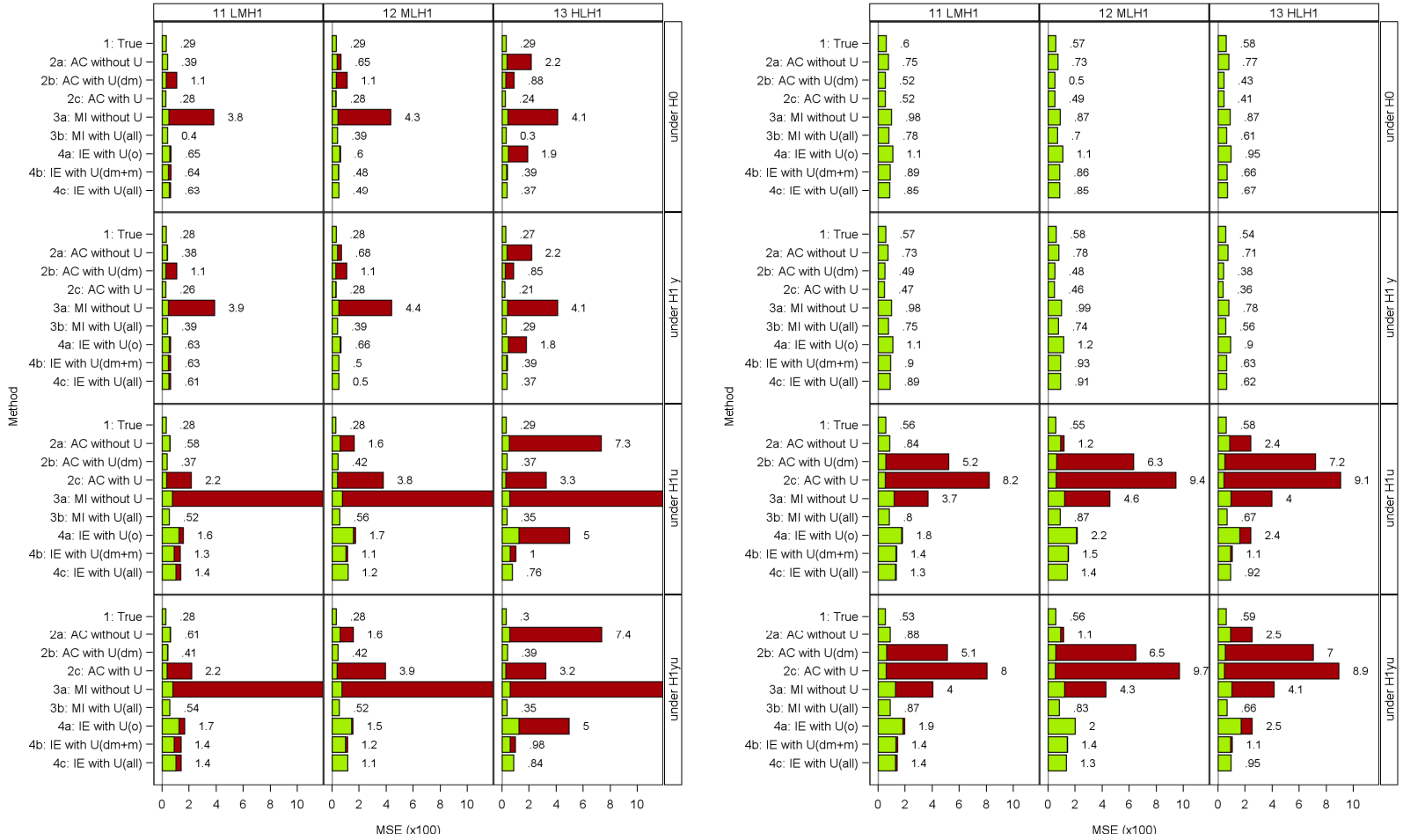
Question 3. Is the degree of association among $\boldsymbol{y}$ *more informative than between* $\boldsymbol{y}$ *and* $\boldsymbol{u}$*?* In Figure 4.8 on page 76, we examined whether high $\boldsymbol{\Sigma}_y$ translates to a correction in bias among 3 levels of $\boldsymbol{\Sigma}_{yu}$. Under a moderate degree of association in $\boldsymbol{\Sigma}_{yu}$, the available case approach (3a) had similar MSE with the IE algorithm approach and there is additional improvement when we adjust for the auxiliary covariate at the time of the event (Method 3b) resulting in MSE that is close to the MI approach. If we are interested in treatment group comparisons under H_{1y}, it seems that method 2b had lower MSE than Method 3b.



(a) MSE for $\widehat{\mu}_{y(A)}$ *at time* $= 4$

(b) MSE for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ *at time* $= 4$

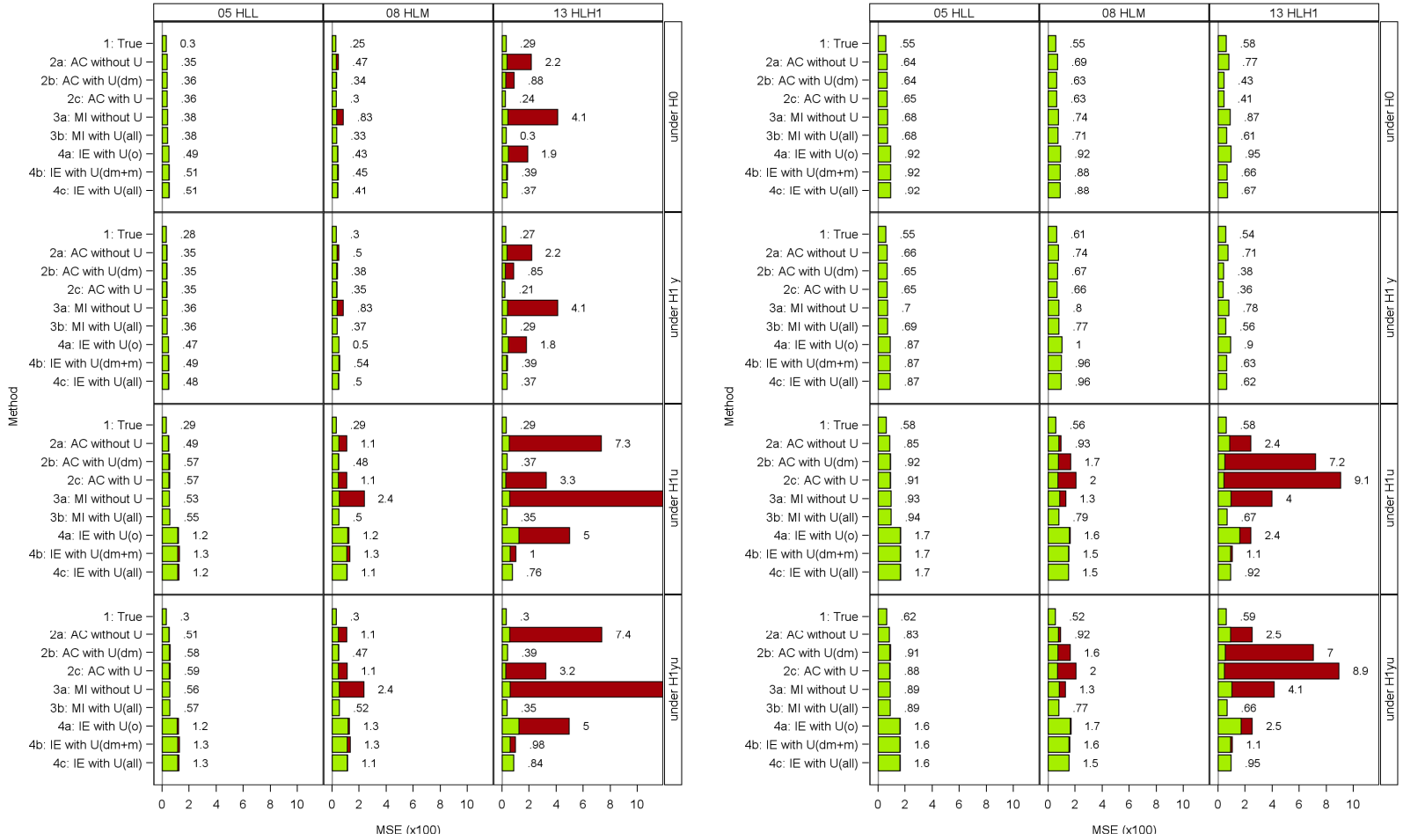
Figure 4.6: Comparison of estimation methods by MSE from selected scenarios to assess the effect of the association between $\boldsymbol{y}$ and $\boldsymbol{u}$ (Q1). The height of the bars represents the MSE (x100) for the parameter which is subdivided into the variance (green) and the square of the bias (red).



(a) MSE for $\widehat{\mu}_{y(A)}$ *at time* $= 4$

(b) MSE for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ *at time* $= 4$

Figure 4.7: Comparison of estimation methods by MSE from selected simulation scenarios to assess whether the degree of association among $\boldsymbol{y}$ more important than the degree of association among $\boldsymbol{u}$ (Q2). The height of the bars represents the MSE (x100) for the parameter which is subdivided into the variance (green) and the square of the bias (red).



(a) MSE for $\widehat{\mu}_{y(A)}$ *at time* $= 4$

(b) MSE for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ *at time* $= 4$

Figure 4.8: Comparison of estimation methods by MSE from selected simulation scenarios to assess whether the degree of association among $\boldsymbol{y}$ more important than the degree of association between $\boldsymbol{u}$ and $\boldsymbol{y}$ (Q3). The height of the bars represents the MSE (x100) for the parameter which is subdivided into the variance (green) and the square of the bias (red).

Question 4. What is the role of the rate of missing?

We are also interested in assessing how much bias does the rate of missing introduce. The missing by design feature of our simulations produce monotone missing so that the rate of missing increases over time and is affected by the rate at which the event occurs in both treatment groups. In Figure 4.9 on page 79, we show the estimated means over time using the 9 estimation methods considered. We see that the bias increases over time due to the increasing rate of missing. The multiple imputation method using the auxiliary information (3b) was able to correct for the bias for all time points. The monotone approach that uses the auxiliary information provides severely biased estimates.

4.5.5 Additional Questions

Question 5. What if we use the conditional mean from the random effects model as the imputed value?

We also considered how the IE algorithm compare to a simple imputation from a linear mixed effects model. In Figure 4.10 on page 80, we summarize the performance metrics for covariance scenario *15 HHH2* for selected estimation methods. The EBLUP imputation method produces estimates that are similar to multiple imputation for $\widehat{\mu}_{y(A)}$. When there were differences in the distribution of the auxiliary covariate between the 2 groups (H_{1y} and H_{1u}), the IE algorithm produced results with less bias. The EBLUP imputation is a good alternative but we were only able to obtain convergent estimates in ninety percent of the simulations.

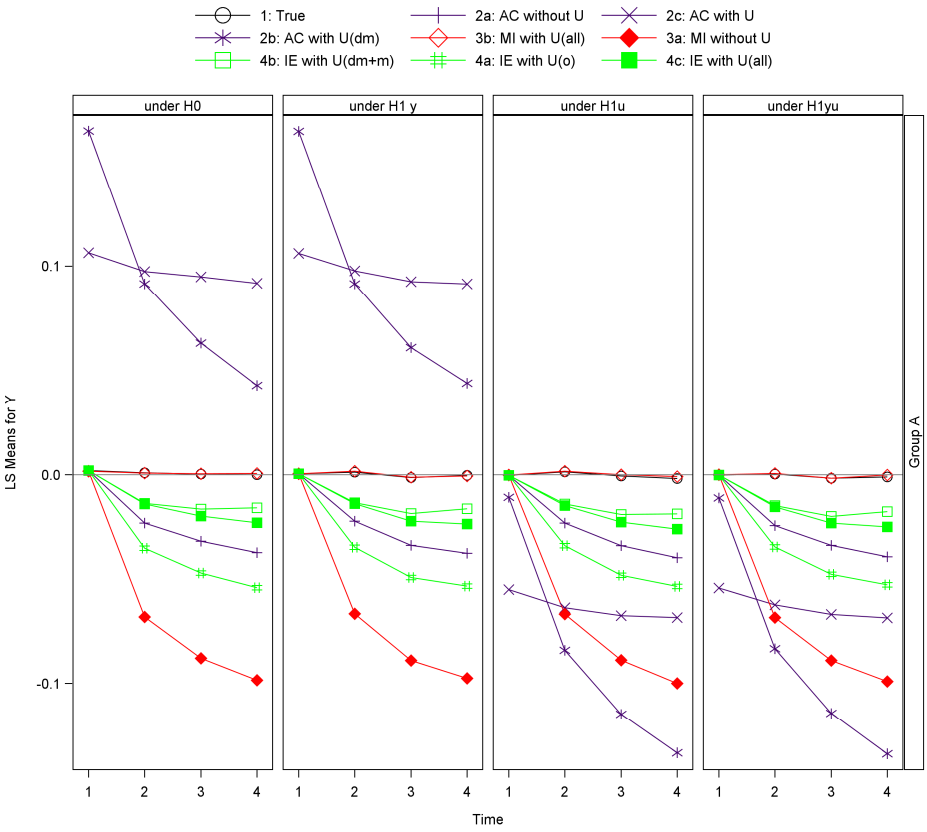
Question 6. What if we apply multiple imputation methods to the IE algorithm?

In some of the scenarios, the IE algorithm improved bias but with poor coverage. We considered how the IE algorithm can be improved by having 5 sets of imputed data sets instead of one to account for the underestimation of the variance. In Figure 4.11 on page 81, we summarize the performance metrics for covariance scenario *12 MLH1* for selected estimation methods. As seen in question 5, EBLUP imputation method produces estimates that are similar to multiple imputation for $\widehat{\mu}_{y(A)}$. When there were differences in

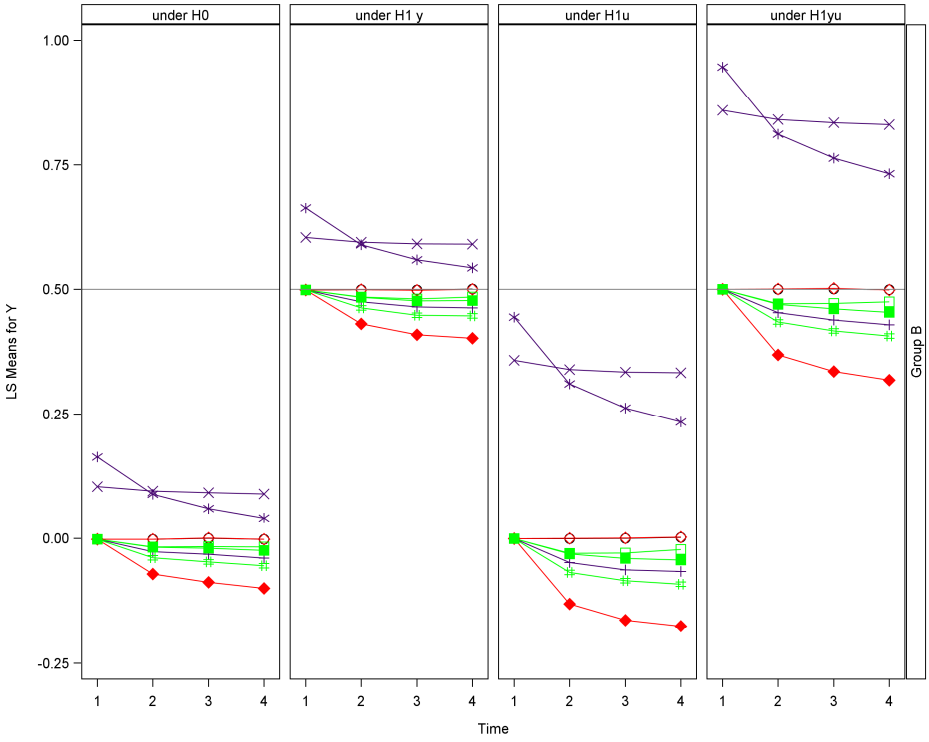
the distribution of the auxiliary covariate between the 2 groups (H_{1y} and H_{1u}), the IE algorithm produced results with less bias. The EBLUP imputation is a good alternative but we were only able to obtain convergent estimates in ninety percent of the simulations. IEMI algorithm improved coverage and MSE for the estimates under all hypotheses and for both parameters, $\widehat{\mu}_{y(A)}$ and $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ at time =4. The performance of IEMI is comparable to multiple imputation but not as conservative (i.e., coverage exceeds 0.95 under all hypotheses).

4.6 Summary

We have shown that use of auxiliary covariates may recover lost information from studies where data are missing by design. Multiple imputation that accounted for the time varying auxiliary covariate consistently corrected the bias, had low MSE and the best coverage among all of the estimation methods considered. However, under MAR, the available case analysis had lower MSE compared to multiple imputation. For the IE algorithm, we observed that the use of the missing partition of the auxiliary covariate vector improved bias especially when the association among the outcomes measurements were strong. Consideration should also be made to whether one is considering a one sample versus a 2 sample problem since some of the estimation differs in performance according to the parameter of interest.

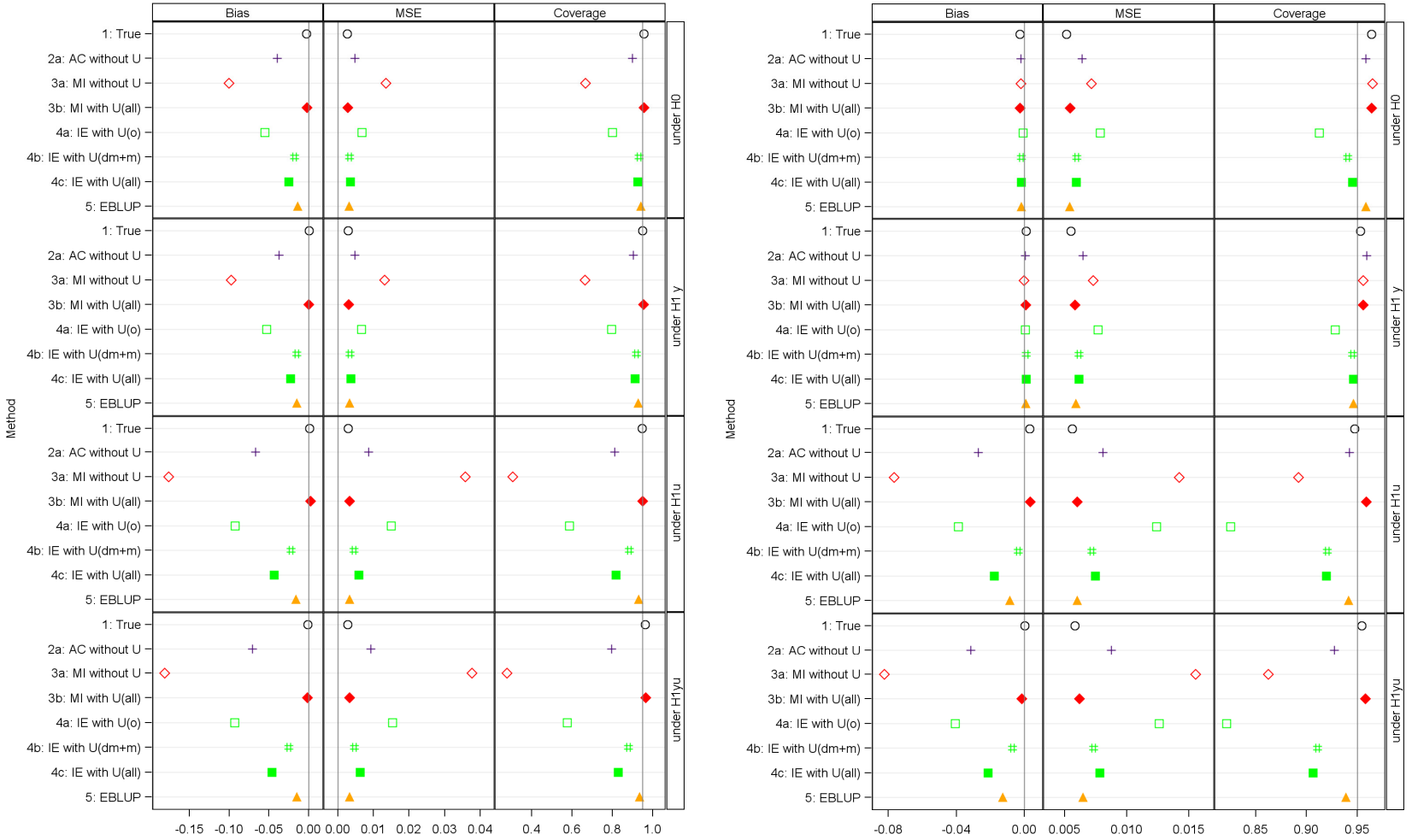


(a) Estimated means for Group A



(b) Estimated means for Group B

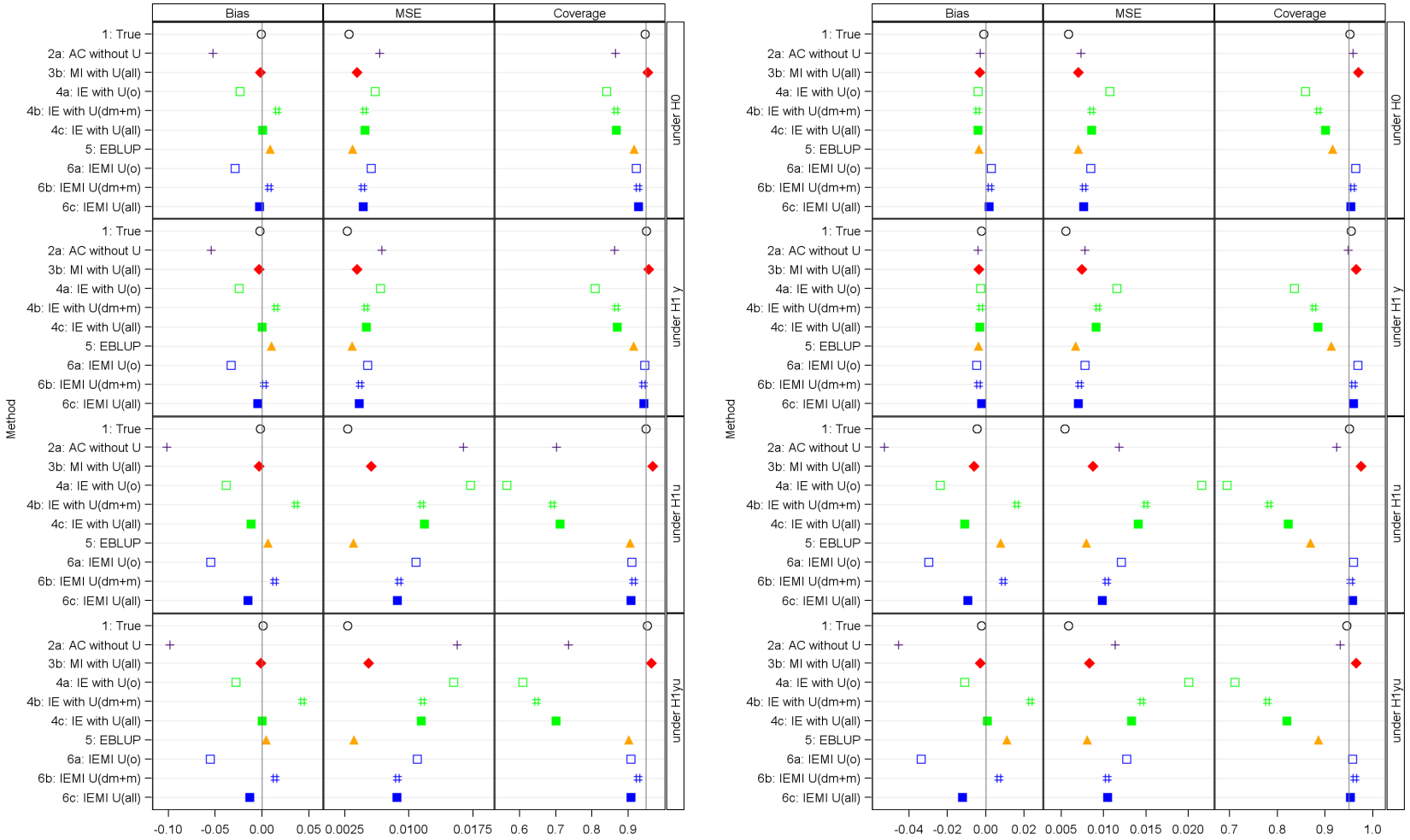
Figure 4.9: Estimated means for covariance scenario 15 $HHH_{(2)}$ over time by group



(a) Performance metrics for $\widehat{\mu}_{y(A)}$ *at time* $= 4$

(b) Performance metrics for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ *at time* $= 4$

Figure 4.10: Comparison of estimation methods with EBLUP imputation by for covariance structure $15HHH_2$ under different hypotheses.



(a) Performance metrics for $\widehat{\mu}_{y(A)}$ *at time* $= 4$

(b) Performance metrics for $\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$ *at time* $= 4$

Figure 4.11: Comparison of estimation methods with IEMI by for covariance structure $12MLH_1$ under different hypotheses.

Chapter 5

APPLICATION TO DPP DATA

We now apply the IE algorithm approach to the missing data from the Diabetes Prevention Program. We obtained the deidentified public release version of the Diabetes Prevention Program datasets from the NIDDK repository based on the data collected before July 2001. The data are consistent but do not match published results since only clinics with IRB approval are included in the release.

Table 5.1: Missing OGTT data by treatment group

Year	Placebo (n = 1030)		Lifestyle (n = 1024)	
	No. Visits (% of randomized)	Missing OGTT (% of Visits)	No. Visits (% of randomized)	Missing OGTT (% of Visits)
Baseline	1030(100%)	6 (0.6%)	1024 (100%)	18 (1.8%)
1	975 (95%)	42 (4.3%)	971 (95%)	15 (1.5%)
2	962 (93%)	148 (15%)	946 (92%)	52 (5.5%)
3	620 (60%)	152 (25%)	600 (59%)	70 (12%)
4	224 (22%)	80 (36%)	228 (22%)	48 (21%)

5.1 Missing Data From OGTT

Table 5.1 summarizes the visits available and missing OGTT measurements in the repository dataset. Due to staggered enrollment over 3 years and an average follow-up of 3.2 years, only 22% of the randomized cohort had year 4 visits so we limit our analyses to year 3. By year 3, the rate of missing OGTT was greater in the Placebo group (36%) than

in the Intensive Lifestyle group (21.1%).

When we examine the patterns of missing 30 minute glucose in Figure 5.1 by diabetes status, the missingness by design is apparent by the preponderance of missing 30 minute values after diabetes diagnosis (in red) while the missing data in blue occurring before diabetes diagnosis are assumed to represent Missing at Random.

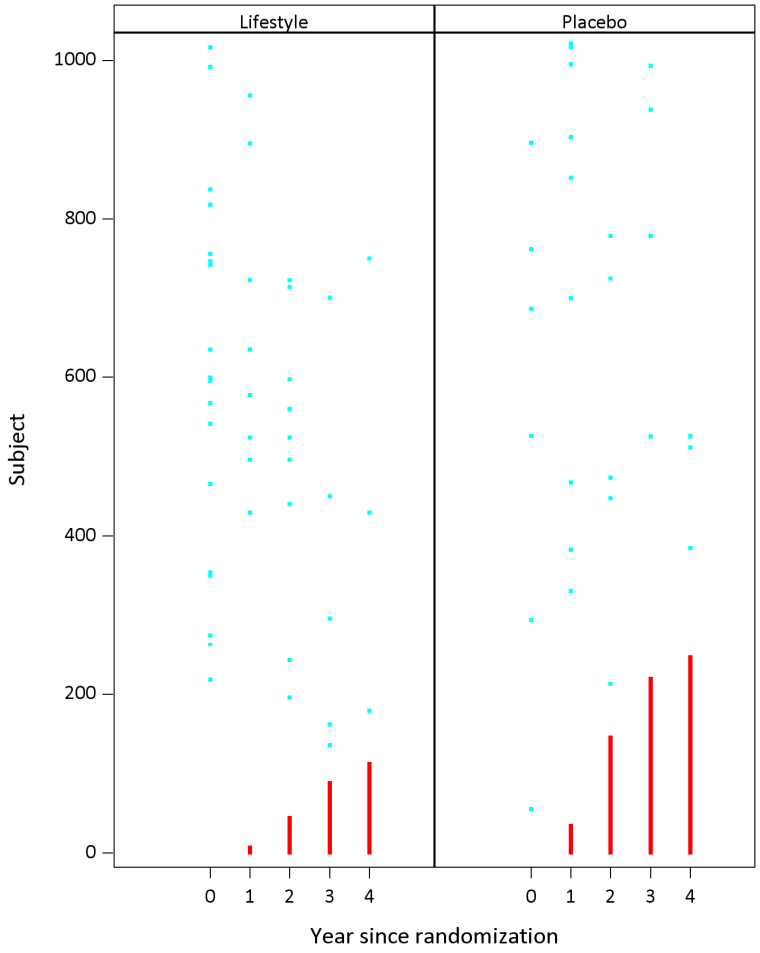


Figure 5.1: Patterns of missing 30 minute glucose (square) by treatment group and diabetes status. Diabetes status is noted as before (blue) and after (red) diabetes diagnosis.

If we replot the first 125 subjects from Figure 5.1 and overlay the availability of the fasting glucose in Figure 5.2 on the next page, we can see that the fasting glucose was observed for a majority of visits with missing 30 minute glucose in years 1-3. Thus, fasting glucose is a good candidate for the time varying auxiliary covariate since missingness on

30 minute glucose depends on the fasting glucose.

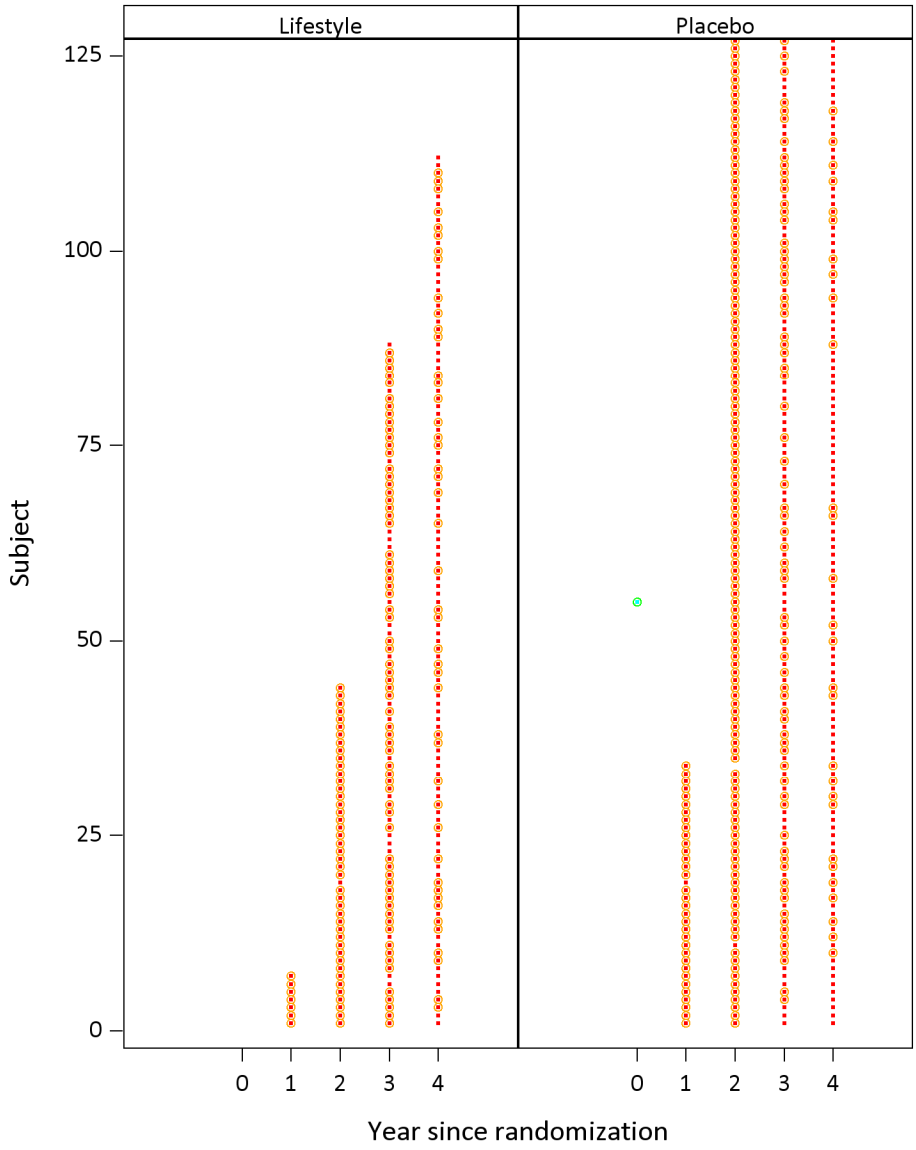


Figure 5.2: Patterns of missing 30 minute glucose (square) and observed fasting glucose (circle) according to treatment group and diabetes status for the first 125 subjects. Diabetes status is noted as before (blue/green) and after (red/orange) diabetes diagnosis.

5.2 Association of Auxiliary Covariate and Outcome

We assume that the association between the 2 measures of glucose do not change after diabetes diagnosis. Table 5.2 on the following page describes the pairwise association

among the 4 annual measures of fasting glucose and the 4 annual measures of 3 minute glucose using pearson correlation. In the upper quadrant of the correlation matrix, fasting glucose and 30 minute glucose taken at the same occasion were strongly correlated and this did not wane over time. The data also shows strong dependence among outcome measures and among auxiliary covariates as seen in the off-diagonal in the upper left and lower right quadrants. From the discussion in 3.2 on page 42, these correlations should yield better prediction of missing outcome values and more efficient inference for unknown model parameters. From our simulation studies, the correlation matrix is similar in magnitude to the covariance scenario 15 $HHH2$ which resulted in good performance for the IE algorithm.

Table 5.2: Pearson correlation matrix of DPP annual fasting (g000) and 30 minute glucose (g030)

Variable	$g000_0$	$g000_1$	$g000_2$	$g000_3$	$g030_0$	$g030_1$	$g030_2$	$g030_1$
$g000_0$	1	0.41	0.46	0.50	0.44	0.25	0.24	0.24
$g000_1$	0.41	1	0.60	0.56	0.20	0.59	0.33	0.30
$g000_2$	0.46	0.60	1	0.63	0.21	0.41	0.56	0.35
$g000_3$	0.50	0.56	0.63	1	0.23	0.38	0.30	0.55
$g030_0$	0.44	0.20	0.21	0.23	1	0.47	0.51	0.52
$g030_1$	0.25	0.59	0.41	0.38	0.47	1	0.56	0.55
$g030_2$	0.24	0.33	0.56	0.30	0.51	0.56	1	0.59
$g030_3$	0.24	0.30	0.35	0.55	0.52	0.55	0.59	1

In Figure 5.3 on the following page, the pairwise associations of the *observed* annual fasting and 30 minute glucose are displayed with the marginal histograms. All tests for normality in the distribution of fasting and 30 minute glucose were significant ($p<0.001$). Jacqmin-Gadda et al. (2007) showed that inference on fixed effects assuming independent gaussian error with homogeneous variance was not impaired when the true error distribution was non-gaussian or heteroscedastic. Thus, our inference on treatment effects are valid without transformations since we are using linear mixed models to estimate the fixed effects.

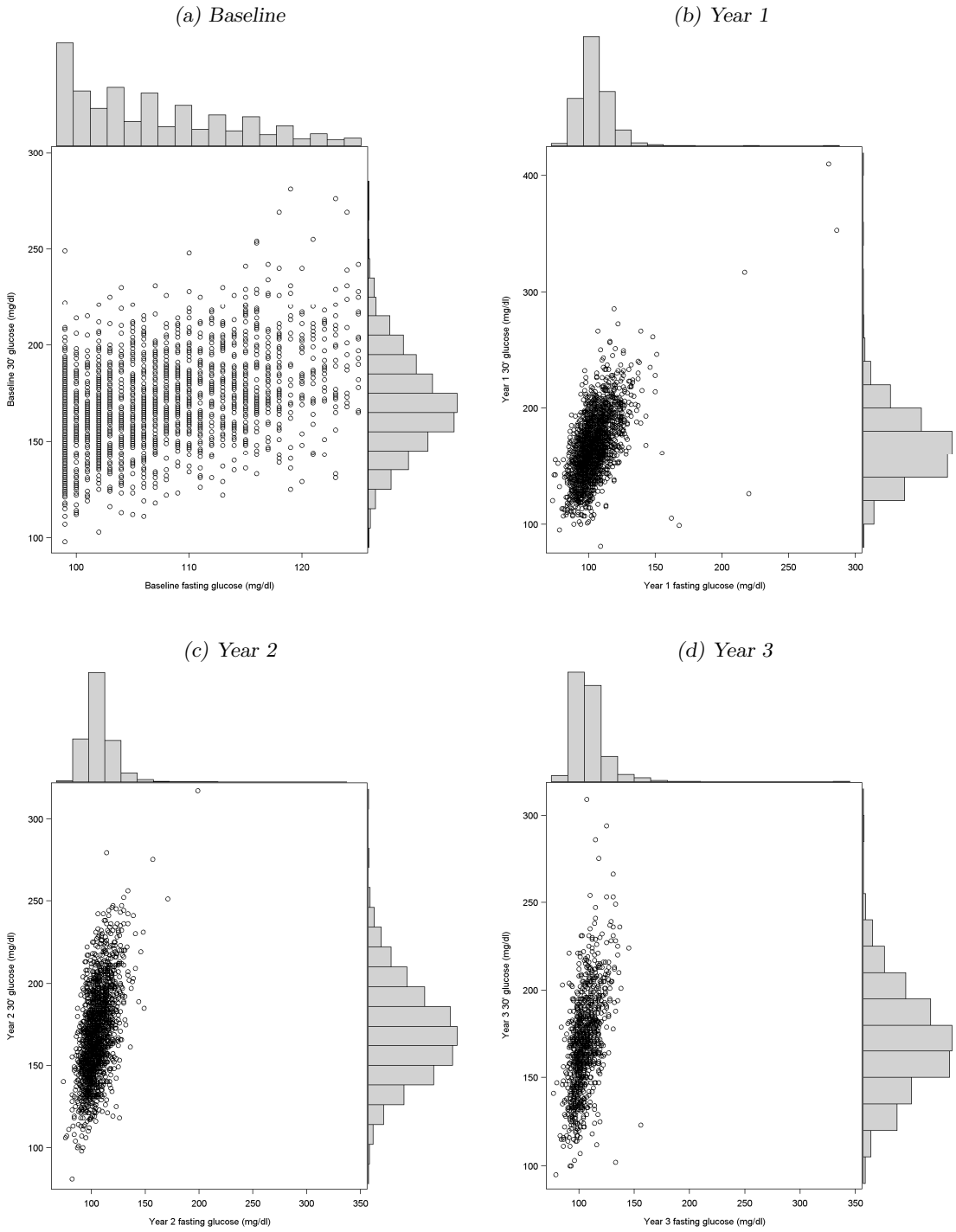


Figure 5.3: Distribution of observed fasting and 30 minute glucose by year

5.3 Mean Estimation

Another way to conduct a conditional mean imputation is to use the estimate for best linear unbiased prediction (EBLUP) from a random effects model as the imputed value for the missing. We considered the performance of this method to the others we considered in the simulation studies.

Figure 5.4 on the next page summarizes the results of applying the methods from the simulations studies plus the EBLUP imputation to the DPP data set. Here, we used the linear mixed effect model with treatment group and time as fixed effects with an interaction term for time $\times$ group. We also assumed an unrestricted covariance structure. The estimated means did not differ among the 6 approaches at baseline and year 1 since the rate of missing was minimal. As the study progressed and more subjects in the Placebo group developed diabetes, the rate of missing increased more in the placebo group than the lifestyle group. This differential rate of missing translated to a bigger difference between the available case and the 2 imputation methods. Because of the strong association between the outcome and auxiliary covariate, the estimated means from multiple imputation (method 3), IE algorithm (method 4c), and IEMI were similar. Among the 3 methods, standard errors were smallest using IE, intermediate in MI, and greatest in IEMI. The worst estimation method is using the available case analysis with adjustment for the auxiliary covariate.

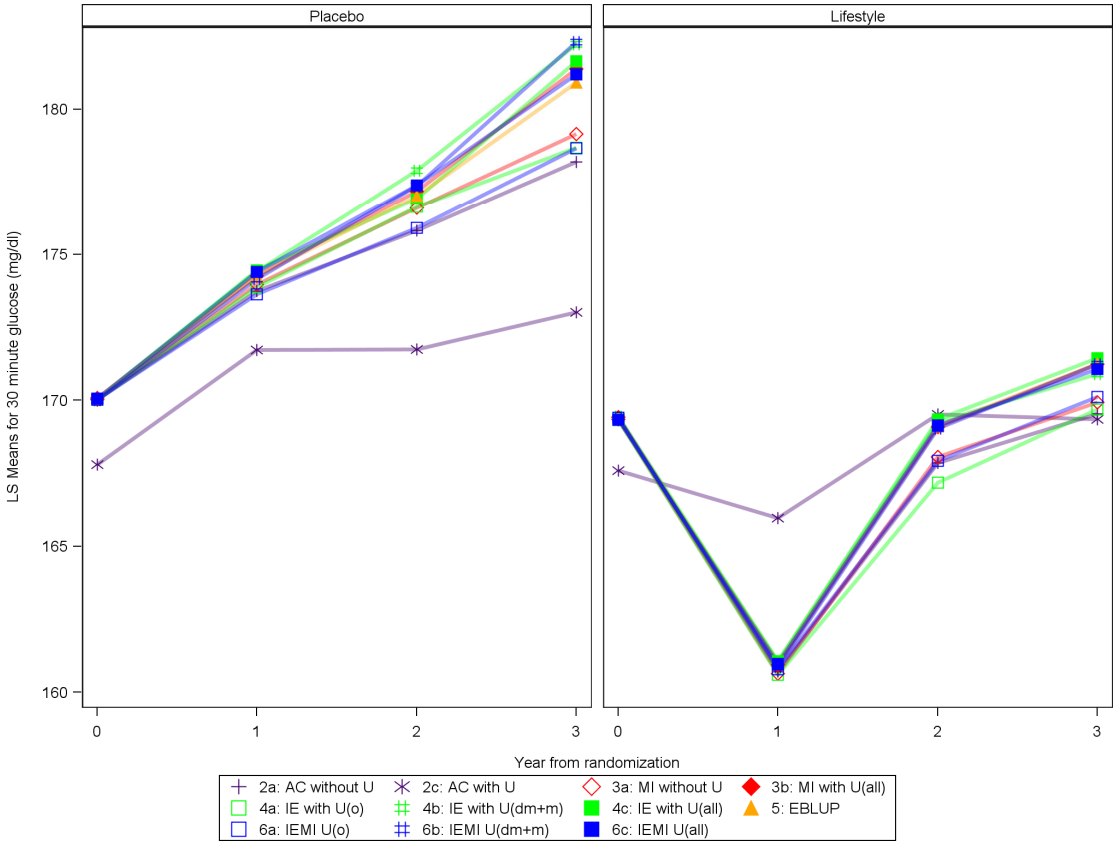


Figure 5.4: Estimated means for 30-minute glucose under different missing data methods

The same data is displayed under Table 5.3 on the following page with p-values for the group differences. The estimated means from the *5: EBLUP* is similar to *3b: MI with U(all)* but with much lower standard errors in the Placebo group at year 3 (1.3 vs 1.04). The standard errors from the IE algorithm using the complete vector is intermediate between the MI and EBLUP approaches. However, all test of treatment differences ledd to the same conclusions with the exception of the naive handling of missing data by *2: AC with U*.

Table 5.3: Estimated means according to treatment group and year using different missing data methods

		Treatment Group				
		Placebo		Lifestyle		
Year	Method	Mean	SE	Mean	SE	P-value
0	2a: AC without U	170.0	0.76	169.4	0.77	0.552
	2c: AC with U	167.8	0.69	167.6	0.69	0.825
	3a: MI without U	170.0	0.77	169.3	0.77	0.536
	3b: MI with U(all)	170.1	0.76	169.4	0.77	0.552
	4a: IE with U(o)	170.0	0.76	169.3	0.76	0.516
	4b: IE with U(dm+m)	170.0	0.76	169.4	0.76	0.583
	4c: IE with U(all)	170.0	0.76	169.4	0.76	0.518
	5: EBLUP	170.0	0.88	169.4	0.88	0.585
	6a: IEMI U(o)	170.0	0.76	169.4	0.76	0.562
	6b: IEMI U(dm+m)	170.0	0.76	169.4	0.76	0.568
	6c: IEMI U(all)	170.1	0.76	169.3	0.76	0.505
1	2a: AC without U	173.7	0.91	160.7	0.91	<.001
	2c: AC with U	171.7	0.75	166.0	0.76	<.001
	3a: MI without U	174.0	0.92	160.6	0.92	<.001
	3b: MI with U(all)	174.3	0.92	160.9	0.92	<.001
	4a: IE with U(o)	173.9	0.90	160.6	0.91	<.001
	4b: IE with U(dm+m)	174.4	0.91	161.0	0.91	<.001
	4c: IE with U(all)	174.5	0.92	161.1	0.92	<.001
	5: EBLUP	174.3	0.90	161.0	0.90	<.001
	6a: IEMI U(o)	173.6	0.91	160.8	0.92	<.001
	6b: IEMI U(dm+m)	174.2	0.91	160.9	0.92	<.001
	6c: IEMI U(all)	174.4	0.93	161.0	0.92	<.001
2	2a: AC without U	175.8	0.95	167.8	0.93	<.001
	2c: AC with U	171.7	0.79	169.5	0.77	0.044
	3a: MI without U	176.6	0.94	168.1	0.96	<.001
	3b: MI with U(all)	177.2	1.02	169.1	1.02	<.001
	4a: IE with U(o)	176.6	0.92	167.2	0.93	<.001
	4b: IE with U(dm+m)	177.9	0.98	169.3	0.99	<.001
	4c: IE with U(all)	176.9	0.98	169.4	0.99	<.001
	5: EBLUP	177.0	0.90	169.1	0.91	<.001
	6a: IEMI U(o)	175.9	1.12	167.9	0.97	<.001
	6b: IEMI U(dm+m)	177.4	1.16	169.0	1.00	<.001
	6c: IEMI U(all)	177.4	1.23	169.1	1.02	<.001
3	2a: AC without U	178.2	1.17	169.5	1.11	<.001
	2c: AC with U	173.0	0.99	169.3	0.93	0.007
	3a: MI without U	179.1	1.31	169.9	1.14	<.001
	3b: MI with U(all)	181.4	1.32	171.2	1.24	<.001
	4a: IE with U(o)	178.7	1.09	169.7	1.11	<.001
	4b: IE with U(dm+m)	182.2	1.17	170.9	1.19	<.001
	4c: IE with U(all)	181.6	1.24	171.5	1.26	<.001
	5: EBLUP	180.9	1.04	171.3	1.05	<.001
	6a: IEMI U(o)	178.7	1.23	170.1	1.29	<.001
	6b: IEMI U(dm+m)	182.3	1.48	171.2	1.23	<.001
	6c: IEMI U(all)	181.2	1.62	171.1	1.35	<.001

5.4 Summary

For the DPP data set where we found that there is strong association between the 30 minute glucose and fasting glucose, the IE algorithm yielded estimates that are similar to the multiple imputation approach but the standard errors were higher from MI. Adjusting for 30 minute glucose did not yield good estimates of the mean when we used available case analysis. The imputation strategies seemed to improve the parameters of interest compared to the available case approach where we ignore the informative monotone missing.

Chapter 6

SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

The goal of this research was to assess the performance of an imputation-estimation algorithm compared to other different estimation methods under a missing by design scenario. The primary contributions of this research are as follows:

1. We assessed the effect of the missing values under two estimation problems (i.e. one sample and 2 sample settings) and showed that there is a difference depending on the parameter of interest.

2. We contrasted the effect of using different partitions from the vector of auxiliary covariate and showed that there is significant improvement in performance when we utilize the missing partition of the auxiliary covariate vector in the IE algorithm approach.

3. We found that under A-MAR, multiple imputation consistently outperforms available case analysis and IE algorithm but with higher standard errors.

4. We showed that even if the primary parameter of interest is treatment group

comparison, it is important to assess the difference in the distribution of the auxiliary covariate that may result in informative monotone missing between the groups.

In this dissertation research, we showed that when data are missing by design due to a time varying auxiliary covariate that exceeds a threshold, multiple imputation provides the best bias correction among all the methods considered, albeit a bit conservative. The proposed method, Imputation Estimation (IE), improved bias but with poor coverage. Applying a multiple imputation strategy to IE improved coverage due to added variability from 2 sources (IEMI). When there is strong dependence between and among auxiliary and outcome variables, there is little difference between multiple imputation and IE. However, when within subject correlations are low or moderate in auxiliary or outcome measures, one should consider IEMI for bias correction and improved coverage than IE.

6.2 Possible Data Applications

In long term studies, it is not unusual for methodologies for the measurement to change. For example, a new assay is developed for measure insulin. It is very important to have quality control of this transition to ensure that the planned analyses are not compromised. We can consider the case where the 2 methods will have some overlap so that the same sample is measured under the new and old assay for an interval of time. The IE algorithm can then be used to assess what the future value of the old assay. Of course, it is important that the overlap should include the range of possible values.

6.3 Possible Methodology Directions

The imputation based on time varying covariates when data are missing by design seems to have good potential. We describe future directions for this algorithm:

1. Consider other covariance estimation methods so that we can model more than 2 longitudinal measures.

2. Possibly consider a selection model approach similar to the Heckit models.

3. Bootstrap estimates for the variance might help in the coverage problem we observed in the simulation studies.

4. Consider other estimation methods such as MLE from EM algorithm or robust variance estimation methods.

5. In our example, the auxiliary covariate is related to the missing data mechanism. We can consider the case where another event was the mechanism for missing. The IE algorithm approach will then have to include two time varying auxiliary covariate, one for the missing data mechanism and another for the correlated auxiliary covariate.

6. The IE algorithm assumes continuous variables for both outcome and auxiliary covariate. We can consider how to extend the IE algorithm to allow for generalized linear models.

7. The IE algorithm assumes a linear relationship between auxiliary and outcome variables, we can consider how nonlinearity can be accommodated.

8. The current version of the IE algorithm assumes that the distribution before and after the event are the same. We can consider a mixture of distribution for the outcome and auxiliary covariates to allow for the difference in means for example.

Bibliography

G.B. Airy. *On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations*. Macmillan, London, 1861.

K.G. Alberti and P.Z. Zimmet. Definition, diagnosis and classification of diabetes mellitus and its complications. part 1. diagnosis and classification of diabetes mellitus provisional report of a who consultation. *Diabetic Medicine*, 15:539 553, 1998.

American Diabetes Association. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, 20:1183–97, 1997.

S.G. Baker and N.M. Laird. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83(401):62–69, 1988.

S.F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B*, 22(2): 302 – 306, 1960.

L.M. Collins, J.L. Schafer, and C. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330, 2001.

N.R. Cook. An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine*, 16(23):2713–2728, 1997.

N.R. Cook. Imputation strategies for blood pressure data nonignorably missing due to medication use. *Clinical Trials*, 3(5):411–20, 2006.

M.J. Daniels and J.W. Hogan. *Missing data in Longitudinal Studies: strategies for Bayesian modeling and sensitivity analysis*, volume 109 of *Monographs on Statistics and Applied Probability*. Chapman and Hall CRC, Boca Raton, 2008.

M.J. Daniels and M. Pourahmadi. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566, 2002.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

DPP Research Group. The diabetes prevention program: Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care*, 22(4):623–34, 1999.

DPP Research Group. Role of insulin secretion and sensitivity in the evolution of type 2 diabetes in the diabetes prevention program: effects of lifestyle intervention and metformin. *Diabetes*, 54(8):2404–14, 2005.

DPP Research Group. 10-year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *Lancet*, 374(9702):1677, 2009.

G.M. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, editors. *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. CRC Press, New York, 2009.

W.D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747, 1991.

A.T. Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics - Theory and Methods*, 23(11): 3105 – 3119, 1994.

A.S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369, 1962.

J. Goodnight. Computing mivque0 estimates of variance components. *SAS Institute*, (R-105), 1978.

J.W. Hardin. The sandwich estimate of variance. In Thomas B. Fomby and R. Carter Hill, editors, *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, volume 17 of *Advances in Econometrics*, pages 45–73. Emerald Group Publishing Limited, 2003.

D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320, 1977.

J.J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492, 1976.

C.R. Henderson. *Applications of Linear Models in Animal Breeding.* University of Guelph, Guelph, Canada, 1984.

P.J. Huber. The behavior of maximum likelihood estimation under nonstandard conditions. In L.M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press, 1967.

J. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *TEST*, 18(1):1, 2009.

H. Jacqmin-Gadda, S. Sibillot, C. Proust, J. Molina, and R. Thiebaut. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis*, 51(10):51425154, Jun 2007.

J.M. Lachin. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21(3):167, 2000.

N.M. Laird and J.H. Ware. Random effects models for longitudinal data: an overview of recent results. *Biometrics*, 38:963–974, 1982.

K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

X. Lin, L. Ryan, M. Sammel, D. Zhang, C. Padungtod, and X. Xu. A scaled linear mixed model for multiple outcomes. *Biometrics*, 56(2):593–601, 2000.

R.J.A. Little. Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. J. Wiley and Sons, New York, 2002.

D.R. Matthews, J.P. Hosker, A.S. Rudenski, B.A. Naylor, D.F. Treacher, and R.C. Turner. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28:412–419, 1985.

W. Navidi. A graphical illustration of the em algorithm. *The American Statistician*, 51 (1):29, 1997.

D.I. Phillips, P.M. Clark, C.N. Hales, and C. Osmond. Understanding oral glucose tolerance: comparison of glucose or insulin measurements during the oral glucose tolerance test with specific measurements of insulin resistance and insulin secretion. *Diabetic Medicine*, 11:286292, 1994.

J. Roy, X. Lin, and L.M. Ryan. Scaled marginal models for multiple continuous outcomes. *Biostatistics*, 4(3):371–383, 2003.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581592, 1976.

D.B. Rubin. Formalizing subjective notions about the effecr if nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72:538–543, 1977.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley and Sons, New York, 1987.

J.L. Schafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177, 2002.

J.L. Schafer and Schenker N. Inference with imputed conditional means. *Journal of the American Statistical Association*, 95(449):144–154, 2000.

K. Tanabe and M. Sagae. An exact cholesky decomposition and the generalized inverse of the variance-covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):211–219, 1992.

R. Thiebaut, H. Jacqmin-Gadda, G. Chene, C. Leport, and D. Commenges. Bivariate linear mixed models using sas proc mixed. *Computer Methods Programs in Biomedicine*, 69(3):249–56, 2002.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag New York, Incorporated, Secaucus, NJ, USA, 2000.

T.M. Wallace and D.R. Matthews. The assessment of insulin resistance in man. *Diabetic Medicine*, 19(7):527–534, 2002.

C. Wang and C.B. Hall. Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*, 29(6):671–9, 2010.

B.T. West, K.B. Welch, and A.T. Galecki. *Linear mixed models: a practical guide using statistical software*. Chapman Hall/CRC, 2007.

R.D. Wolfinger. Covariance structure selection in general mixed models. *Communications in Statistics - Simulation and Computation*, 22(4):1079–1106, 1993.

R.D. Wolfinger. Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(2):205, 1996.

L. Wu. *Mixed-Effects Models with Incomplete Data*. CRC Press, Boca Raton, FL, 2010.

M.C. Wu and K.R. Bailey. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45(3):939–55, 1989.

M.C. Wu and R.J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175, 1988.

L. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in Medicine*, 11:769–782, 1992.

Appendix A

SIMULATION STUDY DETAILS

A.1 Summary of scenarios

Table A.1: Mean number of iterations for EM convergence from 1000 simulation runs by scenario. The covariance are expressed in the notation listed in Table 4.2 on page 59

$\boldsymbol{\Sigma}_y$	$\boldsymbol{\Sigma}_u$	$\boldsymbol{\Sigma}_{yu}$	Scenario	under H_0	under H_{1y}	under H_{1u}	under H_{1yu}
cs(1,0.05,4)	cs(1,0.05,4)	cs(-0.05,-0.05,4)	01 LLL	2	3	2	3
		cs(-0.25,-0.05,4)	06 LLM	3	4	3	4
		cs(-0.6,-0.05,4)	10 LLH1	4	5	4	5
	cs(1,0.25,4)	cs(-0.05,-0.05,4)	02 LML	2	2	2	2
		cs(-0.6,-0.05,4)	11 LMH1	4	5	4	5
cs(1,0.25,4)	cs(1,0.05,4)	cs(-0.05,-0.05,4)	03 MLL	2	3	2	3
		cs(-0.6,-0.05,4)	12 MLH1	4	5	4	5
	cs(1,0.25,4)	cs(-0.05,-0.05,4)	04 MML	2	2	2	2
		cs(-0.25,-0.05,4)	07 MMM	3	4	3	4
		cs(-0.6,-0.25,4)	14 MMH2	4	6	4	6
cs(1,0.6,4)	cs(1,0.05,4)	cs(-0.05,-0.05,4)	05 HLL	2	3	2	3
		cs(-0.25,-0.05,4)	08 HLM	4	6	4	6
		cs(-0.6,-0.05,4)	13 HLH1	7	9	7	9
	cs(1,0.6,4)	cs(-0.25,-0.05,4)	09 HHM	4	4	4	4
		cs(-0.6,-0.25,4)	15 HHH2	5	6	5	6

A.2 Summary of results

Table A.2: Performance of estimation methods under different scenarios for over 1000 simulations with missing rate of 10 percent per year

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c:AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
LLL	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.95	0.95	0.96	0.97	0.82	0.82	0.82
			MSE	0.56	0.78	0.78	0.77	0.94	0.91	1.15	1.14	1.12
			Bias	0.002	0.000	0.001	0.000	0.001	0.001	0.003	0.003	0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.94	0.93	0.97	0.97	0.81	0.81	0.82
			MSE	0.28	0.45	0.41	0.45	0.47	0.46	0.64	0.63	0.62
			Bias	0.001	0.026	0.015	0.026	0.001	0.001	0.025	0.025	0.025
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.93	0.94	0.94	0.94	0.96	0.96	0.81	0.81	0.82
			MSE	0.62	0.86	0.86	0.86	1.03	1.03	1.26	1.26	1.25
			Bias	−0.001	−0.001	−0.001	−0.001	−0.003	−0.005	−0.001	−0.001	−0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.92	0.93	0.91	0.96	0.96	0.79	0.80	0.80
			MSE	0.31	0.49	0.44	0.50	0.50	0.51	0.69	0.69	0.68
			Bias	0.002	0.028	0.017	0.029	0.000	−0.001	0.029	0.029	0.029
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.94	0.94	0.97	0.97	0.68	0.68	0.69
			MSE	0.59	1.05	1.22	1.21	1.29	1.33	2.11	2.10	2.08
			Bias	0.001	0.021	0.045	0.043	−0.000	0.001	0.017	0.018	0.018
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.90	0.90	0.87	0.96	0.96	0.58	0.58	0.59
			MSE	0.29	0.83	0.82	0.99	0.81	0.87	1.73	1.72	1.71
			Bias	0.002	0.050	0.049	0.063	0.003	0.003	0.047	0.048	0.047
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.92	0.92	0.97	0.97	0.74	0.73	0.74
			MSE	0.57	1.08	1.25	1.25	1.32	1.31	1.94	1.93	1.92
			Bias	−0.001	0.022	0.046	0.045	−0.006	−0.007	0.023	0.023	0.024
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.89	0.89	0.86	0.96	0.97	0.60	0.60	0.60
			MSE	0.28	0.84	0.83	1.01	0.84	0.85	1.66	1.65	1.65
			Bias	0.001	0.049	0.048	0.063	−0.004	−0.005	0.050	0.051	0.051
LLM	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.95	0.95	0.96	0.96	0.83	0.85	0.85
			MSE	0.57	0.76	0.73	0.72	0.94	0.92	1.18	1.13	1.08
			Bias	−0.000	−0.001	−0.001	−0.002	−0.002	−0.002	−0.004	−0.004	−0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.93	0.94	0.87	0.97	0.82	0.83	0.84
			MSE	0.28	0.42	0.42	0.41	0.98	0.46	0.60	0.59	0.57

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			Bias	0.000	0.020	−0.022	0.023	−0.073	−0.002	0.018	0.017	0.017
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.94	0.95	0.96	0.97	0.84	0.85	0.85
			MSE	0.63	0.82	0.78	0.78	0.96	0.94	1.16	1.11	1.11
			Bias	0.002	0.005	0.005	0.005	0.005	0.003	0.008	0.008	0.008
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.94	0.92	0.88	0.96	0.80	0.81	0.82
			MSE	0.30	0.48	0.42	0.48	0.92	0.52	0.66	0.63	0.62
			Bias	0.003	0.026	−0.016	0.029	−0.065	0.004	0.030	0.029	0.029
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.75	0.68	0.95	0.96	0.66	0.68	0.70
			MSE	0.64	1.09	2.40	3.12	1.74	1.29	2.31	2.19	2.11
			Bias	0.003	0.023	0.119	0.146	−0.061	0.002	0.026	0.027	0.024
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.90	0.90	0.63	0.77	0.96	0.59	0.61	0.62
			MSE	0.31	0.83	0.76	1.95	2.55	0.87	1.85	1.76	1.68
			Bias	0.003	0.045	0.041	0.116	−0.130	0.003	0.048	0.048	0.045
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.78	0.67	0.94	0.97	0.69	0.70	0.72
			MSE	0.58	1.03	2.25	2.95	1.76	1.27	2.18	2.09	2.02
			Bias	−0.001	0.018	0.113	0.141	−0.065	0.001	0.021	0.023	0.019
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.92	0.92	0.70	0.76	0.96	0.59	0.61	0.65
			MSE	0.29	0.77	0.70	1.79	2.70	0.83	1.66	1.59	1.51
			Bias	0.001	0.039	0.035	0.110	−0.136	−0.000	0.041	0.042	0.038
LLH1	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.95	0.96	0.96	0.97	0.97	0.85	0.89	0.89
			MSE	0.54	0.74	0.48	0.46	0.96	0.75	1.05	0.86	0.85
			Bias	0.001	0.001	0.000	−0.000	0.002	0.001	0.003	0.002	0.002
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.59	0.93	0.31	0.97	0.84	0.87	0.89
			MSE	0.28	0.40	1.07	0.30	4.21	0.40	0.58	0.47	0.46
			Bias	0.002	0.014	−0.089	0.022	−0.193	0.001	0.020	0.016	0.015
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.95	0.96	0.96	0.97	0.96	0.84	0.87	0.87
			MSE	0.57	0.79	0.49	0.47	1.04	0.83	1.14	0.93	0.93
			Bias	−0.001	−0.001	−0.002	−0.001	−0.001	−0.001	0.000	−0.000	−0.000
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.97	0.94	0.57	0.93	0.30	0.96	0.82	0.87	0.88
			MSE	0.26	0.40	1.09	0.28	4.31	0.40	0.59	0.47	0.46
			Bias	0.000	0.013	−0.091	0.019	−0.195	−0.002	0.017	0.012	0.012

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.17	0.03	0.74	0.97	0.70	0.78	0.80
			MSE	0.59	1.06	6.51	10.11	4.66	1.07	2.21	1.51	1.44
			Bias	−0.002	0.009	0.241	0.308	−0.180	−0.001	0.014	0.014	0.012
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.93	0.14	0.05	0.95	0.61	0.69	0.73
			MSE	0.28	0.69	0.46	4.12	14.88	0.71	1.70	1.13	1.07
			Bias	−0.001	0.022	0.014	0.193	−0.374	−0.002	0.028	0.025	0.022
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.14	0.02	0.76	0.98	0.70	0.78	0.80
			MSE	0.59	0.97	6.62	10.35	4.18	0.97	2.02	1.48	1.45
			Bias	0.004	0.014	0.245	0.312	−0.171	0.005	0.016	0.018	0.013
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.93	0.94	0.12	0.06	0.97	0.60	0.67	0.71
			MSE	0.29	0.65	0.43	4.20	14.34	0.64	1.57	1.09	1.03
			Bias	0.001	0.023	0.016	0.195	−0.368	0.002	0.027	0.026	0.022
LML	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.97	0.98	0.85	0.85	0.85
			MSE	0.60	0.78	0.78	0.78	0.91	0.89	1.09	1.09	1.07
			Bias	−0.003	−0.002	−0.002	−0.002	−0.002	−0.002	−0.005	−0.005	−0.005
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.95	0.94	0.97	0.97	0.85	0.85	0.86
			MSE	0.28	0.43	0.40	0.42	0.45	0.46	0.54	0.53	0.53
			Bias	−0.000	0.023	0.015	0.022	0.000	0.000	0.016	0.015	0.015
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.94	0.95	0.97	0.97	0.85	0.85	0.85
			MSE	0.57	0.76	0.77	0.76	0.90	0.89	1.10	1.09	1.09
			Bias	−0.001	0.003	0.003	0.003	0.003	0.003	0.001	0.001	0.002
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.94	0.94	0.97	0.97	0.83	0.83	0.83
			MSE	0.29	0.44	0.41	0.44	0.44	0.45	0.60	0.59	0.59
			Bias	0.000	0.025	0.016	0.024	0.002	0.002	0.020	0.019	0.019
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.91	0.92	0.97	0.96	0.71	0.72	0.72
			MSE	0.56	1.03	1.13	1.12	1.27	1.23	1.93	1.93	1.91
			Bias	−0.002	0.014	0.031	0.031	−0.007	−0.008	0.013	0.013	0.012
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.91	0.91	0.89	0.96	0.96	0.64	0.64	0.65
			MSE	0.28	0.71	0.69	0.78	0.74	0.72	1.40	1.38	1.38
			Bias	0.001	0.041	0.039	0.049	−0.002	−0.003	0.035	0.034	0.033
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.93	0.93	0.97	0.96	0.74	0.74	0.74

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			MSE	0.59	0.97	1.11	1.08	1.22	1.24	1.76	1.75	1.74
			Bias	−0.001	0.022	0.040	0.040	0.004	0.002	0.020	0.021	0.020
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.91	0.92	0.90	0.95	0.96	0.64	0.65	0.65
			MSE	0.30	0.75	0.73	0.82	0.79	0.81	1.38	1.37	1.35
			Bias	0.000	0.044	0.042	0.052	0.004	0.002	0.038	0.037	0.036
LMH1	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.93	0.94	0.94	0.94	0.96	0.96	0.86	0.89	0.90
			MSE	0.60	0.75	0.52	0.52	0.98	0.78	1.07	0.89	0.85
			Bias	−0.002	−0.003	−0.004	−0.004	−0.004	−0.004	−0.004	−0.004	−0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.57	0.94	0.32	0.96	0.81	0.81	0.82
			MSE	0.29	0.39	1.09	0.28	3.81	0.40	0.65	0.64	0.63
			Bias	0.001	0.012	−0.089	−0.007	−0.182	−0.001	−0.031	−0.042	−0.035
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.95	0.96	0.96	0.97	0.86	0.89	0.90
			MSE	0.57	0.73	0.49	0.47	0.98	0.75	1.08	0.90	0.89
			Bias	−0.005	−0.006	−0.003	−0.002	−0.007	−0.005	−0.004	−0.003	−0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.56	0.94	0.33	0.97	0.82	0.81	0.83
			MSE	0.28	0.38	1.08	0.26	3.88	0.39	0.63	0.63	0.61
			Bias	−0.001	0.011	−0.090	−0.007	−0.185	−0.002	−0.031	−0.043	−0.035
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.96	0.20	0.04	0.78	0.97	0.75	0.79	0.81
			MSE	0.56	0.84	5.20	8.19	3.67	0.80	1.80	1.37	1.34
			Bias	−0.001	0.006	0.216	0.277	−0.158	−0.002	−0.027	−0.025	−0.026
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.94	0.36	0.07	0.96	0.61	0.63	0.66
			MSE	0.28	0.58	0.37	2.16	12.28	0.52	1.57	1.34	1.36
			Bias	0.001	0.018	−0.006	0.135	−0.340	−0.002	−0.058	−0.067	−0.059
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.23	0.06	0.77	0.97	0.70	0.79	0.80
			MSE	0.53	0.88	5.11	8.04	4.04	0.87	1.95	1.42	1.40
			Bias	−0.005	0.004	0.212	0.273	−0.167	−0.007	−0.035	−0.033	−0.036
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.94	0.37	0.07	0.96	0.59	0.62	0.63
			MSE	0.28	0.61	0.41	2.17	12.72	0.54	1.66	1.39	1.41
			Bias	−0.001	0.018	−0.006	0.135	−0.345	−0.004	−0.063	−0.071	−0.064
MLL	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.96	0.95	0.95	0.97	0.97	0.85	0.84	0.85
			MSE	0.56	0.74	0.74	0.75	0.84	0.82	1.06	1.06	1.06

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			Bias	−0.007	−0.009	−0.009	−0.008	−0.011	−0.011	−0.010	−0.010	−0.010
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.97	0.97	0.82	0.83	0.83
			MSE	0.30	0.40	0.40	0.41	0.44	0.43	0.57	0.58	0.58
			Bias	−0.004	0.013	0.010	0.016	−0.007	−0.006	0.016	0.018	0.017
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.95	0.96	0.96	0.85	0.85	0.85
			MSE	0.57	0.76	0.76	0.76	0.90	0.89	1.10	1.10	1.10
			Bias	−0.000	0.002	0.002	0.002	0.001	0.001	0.004	0.004	0.004
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.93	0.95	0.96	0.81	0.80	0.81
			MSE	0.30	0.44	0.43	0.45	0.47	0.47	0.64	0.65	0.65
			Bias	0.002	0.021	0.018	0.024	0.003	0.002	0.026	0.028	0.027
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.93	0.93	0.92	0.93	0.96	0.96	0.73	0.73	0.74
			MSE	0.64	1.01	1.13	1.11	1.22	1.18	1.97	1.97	1.95
			Bias	−0.001	0.015	0.038	0.033	0.001	0.000	0.024	0.026	0.023
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.91	0.90	0.94	0.96	0.58	0.57	0.58
			MSE	0.30	0.67	0.74	0.81	0.72	0.70	1.61	1.64	1.61
			Bias	0.000	0.033	0.042	0.048	0.000	−0.002	0.047	0.051	0.048
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.96	0.95	0.69	0.69	0.70
			MSE	0.58	0.99	1.12	1.08	1.19	1.21	2.05	2.04	2.03
			Bias	0.001	0.015	0.038	0.034	0.000	0.000	0.021	0.023	0.021
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.92	0.91	0.90	0.97	0.95	0.58	0.57	0.58
			MSE	0.30	0.70	0.76	0.82	0.72	0.77	1.70	1.73	1.68
			Bias	0.001	0.034	0.042	0.049	0.000	−0.001	0.044	0.048	0.045
MLH1	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.96	0.95	0.94	0.97	0.97	0.86	0.89	0.90
			MSE	0.57	0.73	0.50	0.49	0.87	0.70	1.07	0.86	0.85
			Bias	−0.003	−0.003	−0.005	−0.006	−0.003	−0.003	−0.004	−0.004	−0.004
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.86	0.55	0.92	0.23	0.95	0.84	0.87	0.87
			MSE	0.29	0.65	1.10	0.28	4.32	0.39	0.60	0.48	0.49
			Bias	−0.002	−0.053	−0.090	0.014	−0.197	−0.002	−0.024	0.016	0.000
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.95	0.95	0.97	0.97	0.84	0.88	0.88
			MSE	0.58	0.78	0.48	0.46	0.99	0.74	1.16	0.93	0.91
			Bias	−0.004	−0.004	−0.003	−0.002	−0.005	−0.004	−0.003	−0.003	−0.003

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.86	0.55	0.93	0.23	0.96	0.81	0.87	0.87
			MSE	0.28	0.68	1.07	0.28	4.40	0.39	0.66	0.50	0.50
			Bias	−0.002	−0.054	−0.089	0.015	−0.198	−0.004	−0.025	0.015	−0.000
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.92	0.14	0.03	0.69	0.98	0.69	0.78	0.82
			MSE	0.55	1.18	6.32	9.44	4.57	0.87	2.16	1.50	1.41
			Bias	−0.005	−0.053	0.239	0.297	−0.184	−0.006	−0.024	0.016	−0.011
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.70	0.93	0.14	0.03	0.97	0.56	0.69	0.71
			MSE	0.28	1.63	0.42	3.77	14.82	0.56	1.72	1.15	1.18
			Bias	−0.002	−0.102	0.002	0.184	−0.375	−0.004	−0.038	0.036	−0.012
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.12	0.02	0.71	0.97	0.71	0.78	0.82
			MSE	0.56	1.13	6.46	9.69	4.27	0.83	2.00	1.45	1.33
			Bias	−0.000	−0.045	0.243	0.303	−0.175	−0.003	−0.011	0.023	0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.73	0.93	0.12	0.03	0.96	0.61	0.65	0.70
			MSE	0.28	1.56	0.42	3.95	14.52	0.52	1.52	1.16	1.14
			Bias	0.001	−0.099	0.006	0.189	−0.371	−0.002	−0.028	0.043	−0.000
MML	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.97	0.96	0.85	0.85	0.85
			MSE	0.57	0.78	0.78	0.78	0.88	0.86	1.03	1.03	1.02
			Bias	0.000	−0.000	−0.000	−0.000	−0.000	−0.001	−0.001	−0.001	−0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.94	0.95	0.94	0.97	0.97	0.85	0.85	0.85
			MSE	0.27	0.39	0.38	0.40	0.40	0.41	0.52	0.52	0.52
			Bias	0.001	0.017	0.014	0.019	0.001	0.001	0.016	0.017	0.016
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.95	0.95	0.97	0.97	0.86	0.86	0.86
			MSE	0.58	0.79	0.79	0.79	0.89	0.87	1.05	1.05	1.06
			Bias	−0.003	−0.002	−0.002	−0.002	−0.002	−0.003	−0.005	−0.005	−0.005
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.96	0.96	0.85	0.84	0.85
			MSE	0.29	0.41	0.41	0.42	0.44	0.44	0.54	0.54	0.54
			Bias	−0.000	0.015	0.013	0.017	−0.001	−0.002	0.013	0.014	0.013
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.94	0.94	0.96	0.97	0.75	0.75	0.75
			MSE	0.59	0.89	0.96	0.95	1.08	1.07	1.63	1.62	1.63
			Bias	−0.006	0.007	0.024	0.022	−0.009	−0.009	0.006	0.007	0.006
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.92	0.91	0.90	0.96	0.96	0.66	0.66	0.66

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			MSE	0.29	0.60	0.64	0.69	0.68	0.68	1.18	1.18	1.18
			Bias	−0.001	0.027	0.034	0.038	−0.003	−0.003	0.028	0.030	0.028
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.95	0.93	0.93	0.97	0.97	0.73	0.73	0.73
			MSE	0.55	0.93	1.01	1.00	1.09	1.14	1.73	1.73	1.73
			Bias	−0.003	0.011	0.029	0.027	−0.002	−0.004	0.008	0.009	0.008
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.91	0.91	0.96	0.95	0.66	0.65	0.66
			MSE	0.28	0.61	0.64	0.68	0.68	0.72	1.18	1.18	1.17
			Bias	−0.002	0.026	0.033	0.038	−0.004	−0.005	0.023	0.025	0.023
MMM	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.94	0.94	0.95	0.96	0.86	0.86	0.86
			MSE	0.58	0.76	0.72	0.72	0.86	0.86	1.01	0.99	0.98
			Bias	−0.003	−0.002	−0.002	−0.002	−0.001	−0.002	−0.005	−0.004	−0.004
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.93	0.94	0.85	0.96	0.84	0.84	0.84
			MSE	0.30	0.40	0.41	0.38	0.87	0.43	0.55	0.53	0.53
			Bias	−0.001	−0.006	−0.021	0.010	−0.065	−0.001	−0.009	−0.005	−0.009
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.95	0.95	0.96	0.97	0.85	0.85	0.85
			MSE	0.59	0.76	0.70	0.69	0.85	0.81	1.10	1.06	1.06
			Bias	0.003	0.001	−0.000	−0.000	0.001	0.002	0.001	0.001	0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.96	0.94	0.95	0.87	0.97	0.86	0.87	0.87
			MSE	0.29	0.37	0.39	0.35	0.86	0.42	0.52	0.50	0.50
			Bias	0.002	−0.005	−0.021	0.010	−0.066	−0.000	−0.005	−0.001	−0.004
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.82	0.74	0.94	0.97	0.75	0.76	0.75
			MSE	0.61	0.91	1.80	2.30	1.45	1.07	1.75	1.67	1.68
			Bias	−0.001	−0.008	0.098	0.121	−0.062	−0.003	−0.007	0.003	−0.007
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.94	0.79	0.74	0.94	0.64	0.65	0.66
			MSE	0.30	0.56	0.58	1.16	2.24	0.69	1.27	1.20	1.20
			Bias	0.000	−0.010	0.023	0.080	−0.126	−0.001	−0.010	0.005	−0.009
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.80	0.74	0.93	0.96	0.74	0.75	0.76
			MSE	0.57	0.92	1.82	2.32	1.43	1.03	1.70	1.59	1.59
			Bias	−0.004	−0.007	0.097	0.121	−0.056	−0.002	−0.005	0.004	−0.006
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.93	0.77	0.76	0.96	0.63	0.66	0.66
			MSE	0.30	0.57	0.59	1.16	2.12	0.66	1.23	1.13	1.14

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			Bias	−0.001	−0.009	0.023	0.080	−0.120	0.000	−0.012	0.002	−0.012
MMH2	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.96	0.95	0.94	0.97	0.97	0.85	0.89	0.89
			MSE	0.57	0.72	0.52	0.51	0.88	0.76	1.01	0.85	0.85
			Bias	−0.002	−0.003	−0.004	−0.004	−0.003	−0.004	−0.005	−0.006	−0.005
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.86	0.94	0.59	0.66	0.97	0.70	0.78	0.80
			MSE	0.29	0.62	0.28	0.98	1.79	0.39	0.92	0.70	0.65
			Bias	0.000	0.050	0.007	0.084	−0.117	−0.001	0.062	0.049	0.044
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.95	0.95	0.95	0.96	0.96	0.84	0.87	0.88
			MSE	0.61	0.75	0.49	0.48	0.97	0.83	1.09	0.93	0.92
			Bias	0.002	0.004	0.001	0.001	0.004	0.005	0.004	0.003	0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.84	0.92	0.59	0.65	0.95	0.67	0.75	0.78
			MSE	0.32	0.67	0.29	1.00	1.82	0.44	0.96	0.74	0.68
			Bias	0.000	0.052	0.008	0.085	−0.115	0.001	0.063	0.051	0.046
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.91	0.03	0.01	0.89	0.96	0.70	0.73	0.76
			MSE	0.63	1.10	9.28	11.58	2.14	0.98	2.04	1.70	1.57
			Bias	0.003	0.046	0.295	0.332	−0.098	0.002	0.054	0.056	0.043
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.73	0.17	0.00	0.35	0.96	0.46	0.49	0.58
			MSE	0.29	1.43	3.27	9.05	5.31	0.59	2.46	1.96	1.60
			Bias	0.001	0.096	0.170	0.295	−0.215	−0.001	0.111	0.102	0.084
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.90	0.03	0.01	0.87	0.95	0.73	0.77	0.79
			MSE	0.61	1.11	9.13	11.40	2.26	0.99	1.98	1.61	1.49
			Bias	−0.004	0.042	0.291	0.328	−0.103	−0.004	0.052	0.050	0.038
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.72	0.19	0.00	0.36	0.95	0.42	0.49	0.57
			MSE	0.31	1.53	3.36	9.10	5.39	0.63	2.62	2.03	1.69
			Bias	−0.001	0.096	0.171	0.295	−0.216	−0.002	0.117	0.105	0.087
HLL	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.96	0.96	0.96	0.96	0.97	0.87	0.87	0.87
			MSE	0.55	0.64	0.64	0.65	0.68	0.68	0.92	0.92	0.92
			Bias	0.001	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.95	0.95	0.95	0.96	0.86	0.85	0.86
			MSE	0.30	0.35	0.36	0.36	0.38	0.38	0.49	0.51	0.51
			Bias	0.000	0.007	0.009	0.011	−0.001	−0.001	0.015	0.019	0.017

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.95	0.95	0.95	0.97	0.97	0.88	0.88	0.88
			MSE	0.55	0.66	0.65	0.65	0.70	0.69	0.87	0.87	0.87
			Bias	−0.000	0.001	0.001	0.001	0.002	0.001	0.002	0.002	0.002
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.95	0.95	0.97	0.97	0.86	0.85	0.85
			MSE	0.28	0.35	0.35	0.35	0.36	0.36	0.47	0.49	0.48
			Bias	0.001	0.009	0.011	0.013	0.002	0.001	0.017	0.021	0.020
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.93	0.93	0.97	0.95	0.73	0.73	0.73
			MSE	0.58	0.85	0.92	0.91	0.93	0.94	1.65	1.66	1.66
			Bias	−0.002	0.005	0.022	0.017	−0.004	−0.004	0.011	0.015	0.012
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.92	0.92	0.96	0.96	0.65	0.62	0.64
			MSE	0.29	0.49	0.57	0.57	0.53	0.55	1.20	1.27	1.24
			Bias	−0.000	0.015	0.028	0.027	−0.002	−0.003	0.028	0.036	0.033
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.93	0.93	0.96	0.96	0.73	0.74	0.73
			MSE	0.62	0.83	0.91	0.88	0.89	0.89	1.63	1.64	1.64
			Bias	0.002	0.008	0.025	0.020	0.000	−0.000	0.015	0.019	0.017
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.94	0.92	0.92	0.96	0.95	0.65	0.64	0.64
			MSE	0.30	0.51	0.58	0.59	0.56	0.57	1.21	1.28	1.25
			Bias	−0.000	0.014	0.027	0.026	−0.001	−0.002	0.030	0.038	0.034
HLM	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.96	0.96	0.96	0.96	0.97	0.87	0.87	0.87
			MSE	0.55	0.69	0.63	0.63	0.74	0.71	0.92	0.88	0.88
			Bias	−0.001	0.000	0.000	−0.000	0.001	−0.000	−0.003	−0.003	−0.003
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.90	0.93	0.96	0.83	0.97	0.89	0.88	0.89
			MSE	0.25	0.47	0.34	0.30	0.83	0.33	0.43	0.45	0.41
			Bias	−0.001	−0.041	−0.023	0.011	−0.071	−0.001	−0.018	0.023	0.012
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.95	0.96	0.96	0.95	0.86	0.86	0.86
			MSE	0.61	0.74	0.67	0.66	0.80	0.77	1.00	0.96	0.96
			Bias	−0.002	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.91	0.93	0.94	0.83	0.96	0.86	0.84	0.85
			MSE	0.30	0.50	0.38	0.35	0.83	0.37	0.50	0.54	0.50
			Bias	0.000	−0.039	−0.021	0.013	−0.068	0.001	−0.015	0.026	0.016
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.79	0.74	0.92	0.97	0.73	0.74	0.76

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
			MSE	0.56	0.93	1.65	2.05	1.31	0.79	1.62	1.53	1.49
			Bias	−0.003	−0.040	0.096	0.115	−0.067	−0.004	−0.022	0.021	0.002
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.78	0.94	0.78	0.59	0.96	0.64	0.61	0.65
			MSE	0.29	1.09	0.48	1.08	2.36	0.50	1.23	1.30	1.11
			Bias	0.000	−0.077	0.014	0.080	−0.135	−0.002	−0.032	0.048	0.020
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.93	0.81	0.75	0.92	0.96	0.75	0.75	0.76
			MSE	0.52	0.92	1.63	2.04	1.28	0.77	1.68	1.59	1.53
			Bias	−0.000	−0.041	0.096	0.116	−0.068	−0.001	−0.021	0.023	0.004
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.78	0.95	0.77	0.60	0.96	0.66	0.61	0.66
			MSE	0.30	1.08	0.47	1.11	2.36	0.52	1.27	1.35	1.14
			Bias	0.002	−0.078	0.014	0.081	−0.136	0.001	−0.032	0.050	0.021
HLH1	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.94	0.94	0.95	0.96	0.89	0.92	0.92
			MSE	0.58	0.77	0.43	0.41	0.87	0.61	0.95	0.66	0.67
			Bias	0.001	0.003	0.001	0.001	0.003	0.002	0.001	0.000	0.000
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.40	0.57	0.93	0.18	0.96	0.43	0.91	0.92
			MSE	0.29	2.17	0.88	0.24	4.10	0.30	1.88	0.39	0.37
			Bias	0.001	−0.134	−0.078	0.010	−0.192	0.001	−0.119	0.025	0.000
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.95	0.96	0.96	0.97	0.96	0.89	0.92	0.93
			MSE	0.54	0.71	0.38	0.36	0.78	0.56	0.90	0.63	0.62
			Bias	0.000	0.001	−0.002	−0.003	0.000	−0.000	0.001	−0.000	−0.000
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.38	0.59	0.94	0.16	0.96	0.45	0.90	0.92
			MSE	0.27	2.17	0.85	0.21	4.09	0.29	1.80	0.39	0.37
			Bias	0.001	−0.135	−0.078	0.010	−0.193	0.001	−0.116	0.027	0.003
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.73	0.04	0.00	0.61	0.95	0.67	0.85	0.89
			MSE	0.58	2.40	7.19	9.06	3.99	0.67	2.41	1.05	0.92
			Bias	0.001	−0.124	0.259	0.294	−0.174	0.001	−0.091	0.038	0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.03	0.90	0.05	0.00	0.95	0.22	0.66	0.80
			MSE	0.29	7.32	0.37	3.26	14.02	0.35	4.97	1.00	0.76
			Bias	0.002	−0.261	0.009	0.173	−0.367	0.002	−0.193	0.064	0.002
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.70	0.05	0.01	0.60	0.96	0.65	0.83	0.87
			MSE	0.59	2.50	7.03	8.91	4.15	0.66	2.52	1.05	0.95

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_{o}$	4b: IE $\boldsymbol{u}_{m}$	4c: IE $\boldsymbol{u}_{all}$
			Bias	−0.001	−0.127	0.255	0.291	−0.177	−0.001	−0.091	0.036	0.000
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.04	0.89	0.07	0.00	0.95	0.21	0.67	0.81
			MSE	0.30	7.37	0.39	3.21	14.07	0.35	4.96	0.98	0.84
			Bias	0.002	−0.261	0.007	0.172	−0.367	0.001	−0.193	0.064	0.002
HHM	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.94	0.93	0.94	0.94	0.95	0.94	0.90	0.91	0.91
			MSE	0.61	0.71	0.67	0.67	0.76	0.71	0.84	0.81	0.81
			Bias	0.006	0.005	0.003	0.003	0.004	0.004	0.002	0.001	0.002
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.94	0.92	0.94	0.92	0.87	0.96	0.86	0.90	0.88
			MSE	0.29	0.41	0.32	0.41	0.63	0.34	0.48	0.41	0.44
			Bias	0.004	−0.026	0.009	0.031	−0.053	0.004	−0.030	−0.018	−0.025
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.93	0.94	0.95	0.95	0.96	0.95	0.91	0.91	0.91
			MSE	0.62	0.70	0.63	0.64	0.74	0.69	0.81	0.77	0.77
			Bias	0.004	0.005	0.005	0.004	0.006	0.004	0.006	0.006	0.006
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.92	0.95	0.92	0.88	0.95	0.87	0.89	0.89
			MSE	0.29	0.41	0.31	0.40	0.63	0.34	0.48	0.40	0.44
			Bias	0.003	−0.028	0.008	0.030	−0.053	0.003	−0.029	−0.017	−0.025
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.95	0.50	0.39	0.94	0.97	0.86	0.89	0.88
			MSE	0.56	0.77	3.27	4.10	1.00	0.71	1.01	0.88	0.93
			Bias	0.001	−0.026	0.161	0.185	−0.047	−0.001	−0.026	−0.009	−0.023
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.85	0.73	0.41	0.69	0.96	0.71	0.81	0.75
			MSE	0.29	0.71	1.07	2.15	1.50	0.41	0.97	0.67	0.84
			Bias	0.001	−0.056	0.082	0.132	−0.103	0.000	−0.058	−0.029	−0.051
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.49	0.38	0.94	0.95	0.85	0.86	0.85
			MSE	0.57	0.82	3.35	4.22	1.06	0.75	1.15	1.02	1.05
			Bias	0.002	−0.025	0.163	0.188	−0.047	0.001	−0.021	−0.004	−0.018
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.85	0.74	0.42	0.71	0.95	0.72	0.81	0.76
			MSE	0.28	0.72	1.07	2.16	1.51	0.40	0.96	0.66	0.82
			Bias	0.001	−0.056	0.083	0.134	−0.103	0.001	−0.055	−0.026	−0.047
HHH2	H_0	$\widehat{\mu}_{y(A)}$	Coverage	0.96	0.96	0.95	0.95	0.97	0.96	0.91	0.94	0.95
			MSE	0.51	0.64	0.42	0.41	0.72	0.54	0.79	0.60	0.59
			Bias	−0.003	−0.002	−0.002	−0.002	−0.002	−0.003	−0.001	−0.002	−0.002

Continued on Next Page...

Table A.2 – Continued

Scenario	Hyp	Par	Metric	1: True	2a: AC only	2b: AC $\boldsymbol{u}_{dm}$	2c: AC $\boldsymbol{u}_{all}$	3a: MI no $\boldsymbol{u}$	3b: MI $\boldsymbol{u}$	4a: IE $\boldsymbol{u}_o$	4b: IE $\boldsymbol{u}_m$	4c: IE $\boldsymbol{u}_{all}$
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.90	0.85	0.50	0.67	0.96	0.80	0.93	0.93
			MSE	0.26	0.47	0.40	1.02	1.35	0.28	0.68	0.32	0.35
			Bias	−0.002	−0.040	0.041	0.090	−0.100	−0.002	−0.055	−0.018	−0.025
	H_{1y}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.96	0.95	0.95	0.96	0.96	0.93	0.95	0.95
			MSE	0.55	0.65	0.43	0.42	0.73	0.58	0.77	0.61	0.62
			Bias	0.001	0.001	−0.001	−0.002	−0.000	0.001	0.000	0.001	0.001
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.90	0.84	0.51	0.66	0.95	0.80	0.92	0.91
			MSE	0.28	0.47	0.42	1.03	1.31	0.30	0.67	0.34	0.37
			Bias	0.001	−0.037	0.043	0.090	−0.098	0.000	−0.053	−0.015	−0.023
	H_{1u}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.94	0.00	0.00	0.89	0.96	0.83	0.92	0.92
			MSE	0.56	0.81	13.94	16.54	1.42	0.60	1.24	0.72	0.75
			Bias	0.003	−0.027	0.367	0.402	−0.077	0.003	−0.039	−0.004	−0.018
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.95	0.81	0.01	0.00	0.31	0.95	0.59	0.88	0.82
			MSE	0.29	0.86	5.79	11.35	3.58	0.32	1.49	0.44	0.59
			Bias	0.001	−0.067	0.234	0.333	−0.177	0.002	−0.092	−0.022	−0.044
	H_{1yu}	$\widehat{\mu}_{y(A)}$	Coverage	0.95	0.93	0.00	0.00	0.86	0.96	0.82	0.91	0.91
			MSE	0.58	0.88	13.85	16.43	1.55	0.62	1.26	0.73	0.78
			Bias	0.000	−0.032	0.366	0.400	−0.082	−0.002	−0.041	−0.007	−0.021
		$\widehat{\mu}_{y(B)} - \widehat{\mu}_{y(A)}$	Coverage	0.96	0.80	0.01	0.00	0.28	0.96	0.58	0.88	0.83
			MSE	0.28	0.92	5.68	11.20	3.76	0.32	1.53	0.46	0.62
			Bias	−0.001	−0.071	0.232	0.331	−0.181	−0.002	−0.093	−0.025	−0.046