INTERMITTENT DATA IN MARGINAL STRUCTURAL MODELS

by

VANESSA RICHARDSON

B.A., Colorado College, 2008

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Biostatistics and Informatics

2020
This thesis for the Master of Science degree by

Vanessa Richardson

has been approved for the

Department of Biostatistics and Informatics

by

Matthew J. Strand, Chair

Nathan Rabinovitch

Fan Yang

Date: 11 December 2020
Marginal structural models (MSMs) with inverse probability of treatment weighting (IPTW) are commonly used to estimate causal effects in longitudinal observational studies. They have been shown to provide more accurate estimates than traditional methods in the presence of time-varying confounders that are affected by previous treatment. However, the treatment of missing data in these studies is challenging because of the multilevel structure of the data and the way the weights are compounded in the MSM with IPTW framework.

This project applied several methods of handling missing data in MSMs with IPTW to a study of asthmatic children at the Kunsberg School at National Jewish Health where the data were collected intermittently. Methods included multiple imputation on the data; filling in the missing weights with the last value carried forward (LVCF); filling in the weights by restarting them at one each time there is a missing value; and four different ways of filling in gaps in the individual probabilities used to create the weights: (1) average value by subject, (2) average of the two probabilities on either side of the gap, (3) linear interpolation, and (4) the average of randomly generated values. The estimate of interest was the effect of medication use on FEV1 after adjusting for time-varying effects of asthma symptoms. The performance of the different methods was compared using a simulation study.

The simulation results suggested that filling in the weights by restarting them at one at each gap is the most appropriate of those tested for the Kunsberg data. When this method was applied to the Kunsberg data, we obtained an estimate of 0.076 (95% CI: -0.020, 0.171) for the effect of medication use on FEV1. This effect was not statistically significant.
The form and content of this abstract are approved. I recommend its publication.

Approved: Matthew J. Strand
# TABLE OF CONTENTS

**CHAPTER**

I.  INTRODUCTION..................................................................................................................................................1

II.  BASIC THEORY AND METHODS.................................................................................................................2

III. APPLICATION OF A MARGINAL STRUCTURAL MODEL...............................................................................7

IV.  SIMULATION...................................................................................................................................................15

V.  DISCUSSION....................................................................................................................................................22

REFERENCES.........................................................................................................................................................24

APPENDIX.............................................................................................................................................................26
CHAPTER I
INTRODUCTION

Observational data are a key source of insight into medical and biological questions when randomization is difficult or unethical. However, analysis of this kind of data presents special challenges, especially when repeated measures are taken over time. Longitudinal observational studies often feature time-varying covariates; these can be exposures, confounders, or outcomes whose values change over time. Time-varying confounders are variables that are predictors of both the outcome of interest and subsequent exposure. In some cases these confounders are also affected by previous treatment. For example, in a study of the effect of exercise on osteoarthritis outcomes, current knee pain is a time-varying confounder that both affects subsequent exercise and is affected by previous exercise. When a time-varying confounder is affected by previous treatment, standard approaches for adjustment of confounding have been shown to be biased. Marginal structural models (MSMs) are a class of causal models developed by Robins, Hernan, and Brumback to address this problem.

This project will present an application of a repeated measures MSM to a longitudinal observational dataset with a time-varying treatment, a time-varying outcome, and a time-varying confounder that were collected intermittently. It will build upon previous work on this dataset conducted by Lu. Specifically, it will focus on a comparison of different methods for handling intermittent data in marginal structural models. Methods fall into one of two broad categories: (1) imputation of the individual probabilities or the stabilized weights, or (2) imputation of the data. Comparison of the results from each method on the dataset will be supplemented by a simulation study.
CHAPTER II

BASIC THEORY & METHODS

Notation

Notation will follow that set forth in Daniel et. al. A longitudinal setting is one in which \( n \) subjects (\( i = 1, \ldots, n \)) enter a study at time 0 and are subsequently followed until the end of the study period at time \( t (t = 1, \ldots, T) \). The exposure or treatment of interest is represented by \( A \), where \( A_{t,i} \) is the treatment for subject \( i \) at time \( t \). Covariates of interest are labeled \( L \), with \( L_{t,i} \) being covariate value(s) for subject \( i \) at time \( t \). It is assumed that values for \( A_i \) and \( L_i \) remain unchanged during the interval between time \( t \) and time \( t+1 \). The treatment history of subject \( i \) is denoted \( A_{t,i} = (A_{0,i}, \ldots, A_{t,i}) \), and the covariate history of subject \( i \) is denoted \( L_{t,i} = (L_{0,i}, \ldots, L_{t,i}) \). Depending on the study structure and the question of interest, observed outcomes \( Y_i \) can either be measured once at the end of the study or at each time point, with \( Y_{t,i} \) being the outcome for subject \( i \) at time \( t \). Capital letters represent random variables, and lower-case letters represent realizations of random variables.

Potential Outcomes and Causal Effects

An ideal study of the effect of treatment \( A \) on outcome \( Y \) would compare an individual’s outcome given a certain treatment with the outcome that would have been seen for that same individual if, counter to fact, s/he had been given a different treatment. This unseen outcome is referred to as potential or counterfactual. Randomized controlled trials attempt to imitate this type of study by comparing groups that are identical in every way except for their treatment status. Observational studies cannot do the same because treatment status is not assigned. Instead, causal models for observational data rely on a set of assumptions to make inference. The four major assumptions - ignorability, positivity, consistency, and stable unit treatment value - are discussed below. The potential outcome that would have been observed if subject \( i \) had received treatment history \( \alpha \) is denoted \( Y_i(\alpha) \).

When all assumptions are met the distribution of \( Y(\alpha) \) in the population for every realization of
treatment history gives the causal effect of treatment history $\bar{A}$ on outcome $Y$. The aspect of the distribution of $Y(\bar{a})$ that is of interest depends on the outcome and the data structure. For this project we are interested in the mean, $E[Y(\bar{a})]$.

**Time-dependent Confounding**

Time-dependent confounding exists when two conditions are met: (1) each $A_{t,i}$ is causally influenced by treatment at the previous time, $A_{t-1,i}$, and covariates at the current time, $L_{t,i};$ and (2) each $Y_{t,i}$ is causally influenced by both treatment history and covariate history up to time $t$. In a subset of cases involving time-dependent confounding, the confounder $L_{t,i}$ is also causally influenced by previous treatment $A_{t-1,i}$. The confounder is then also on the causal pathway between the treatment and the outcome. Figure II.1 illustrates this type of time-dependent confounding. When this type of time-dependent confounding exists, standard regression methods that control for $L$ can be biased and a marginal structural model is often a more appropriate approach.

![Figure II.1: Directed acyclic graph of the relationship between L, A, and Y](image)

**Marginal Structural Models**

Marginal structural models were designed to address the bias created when standard approaches to confounder adjustment are applied to data with time-varying covariates. The classical MSM considers treatment and confounders that vary over time and a single outcome at the end of the
study period, and thus is defined as a model for the population mean of the counterfactual outcome under treatment history $\overline{A}$,

$$E[Y(\overline{a})] = g((\overline{a}); \gamma),$$

where $g$ is a user-defined function and the $\gamma$ parameters are the causal effect of treatment history on the outcome. Repeated measures MSMs model the mean counterfactual outcome at each time point $t + 1$ as a function of treatment history up to time $t$,

$$E[Y(\overline{a}(t + 1))] = g((\overline{a}(t)); \gamma^6).$$

Both time-invariant and time-varying covariates can be included in the model. There are three ways of estimating coefficients in MSMs: $g$-computation, double robust estimator, and inverse probability of treatment weighting (IPTW). The IPTW method will be used for this project, as it is relatively easy to apply and interpret.

A marginal structural model using IPTW estimation is a multi-step procedure. Step 1 involves the estimation of weights based on inverse probability of treatment. For a binary treatment, two logistic models are fit: one with treatment as the outcome and time-invariant covariates $V$ as predictors, along with treatment history:

$$\text{logit} \ P(A_t = 1|V, \overline{A}_t) = \alpha_0t + \alpha_1tV + \alpha_2t\overline{A}_t,$$

and another with treatment as the outcome and both time-invariant covariates and time-varying confounders as predictors, along with treatment history:

$$\text{logit} \ P(A_t = 1|V, \overline{A}_t, \overline{L}_t) = \alpha'_0t + \alpha'_{1t}V + \alpha'_{2t}\overline{A}_t + \alpha'_{3t}\overline{L}_t.$$

Individual probabilities from these models are then used to create the numerator and denominator of the stabilized weight at each time point for each subject. The numerator of the stabilized weight is the conditional probability of receiving the observed treatment given time-invariant confounders, and the denominator is the conditional probability of receiving the observed treatment.
given both time-invariant and time-varying confounders. The weights are compounded at each time point:

\[ SW_{t+1} = \prod_{i=1}^{t-1} \frac{P(At = 1|V, A_t)}{P(At = 1|V, A_t, L_t)}. \]

The unstabilized version of the weights uses the same denominator as above with a numerator of 1. The unstabilized weights can be used in place of stabilized weights, but they generally provide less efficient estimates.

Step 2 fits a weighted repeated measures model of the outcome of interest on treatment. The weighting creates a pseudo-population with balance in all covariates. Time-varying confounders are not included, because they have been accounted for in the weights. Time-invariant covariates may be included.

Several assumptions are made when fitting a marginal structural model. The Stable Unit Treatment Value Assumption (SUTVA) states both that there is only one version of each treatment and that the treatment assigned to one subject does not affect the potential outcomes for another subject. The positivity assumption states that no covariate can predict treatment status perfectly, i.e. for subjects with each combination of covariates, the probability of treatment is neither zero nor one. The Sequential Ignorability Assumption states that treatment assignment at each time point is independent of future potential outcomes given past observed treatment history, outcome history, and covariates. The consistency assumption states that the potential outcome under each realization of the treatment history equals the observed outcome for that particular treatment history. It is also assumed there is no misclassification bias and no unmeasured confounding.

**Missing Data & Marginal Structural Models**

In any longitudinal study with missing data, care must be taken to choose an appropriate method for handling the missingness. Though it is common for researchers to employ complete case or available case methods in which patients or records with missing data are simply excluded, this
approach may lead to biased estimates or loss of precision. MSMs using IPTW present an additional challenge: because the weights are compounded at each time point, missing data at one time point may affect the values of the weights at all subsequent time points.

Some simulation studies have been conducted to compare methods for addressing missing data in marginal structural models. Moodie et. al. compared multiple imputation (MI) and inverse probability of missingness weighting (IPMW). They recommend MI over IPMW, especially when missing data are strongly predicted by available data. However, they only considered up to 50% missingness in a single time-varying confounder, and their outcome was a single event at the end of the study. Shortreed and Forbes concluded that complete case (CC) analysis showed least bias, but they considered up to 32% missingness only in the exposure. Vourli and Touloumi compared last observation carried forward (LOCF), MI, and IPMW. They recommend IPMW over MI, especially when missing data are strongly predicted by available data. They considered up to 83% missingness, but their missing data were only in a single time-varying confounder and their outcome was a single event at the end of the study. Liu et. al. compared CC, MI, and IPMW. They concluded that MI produced the least biased estimates, but they only considered up to 30% missingness in a time-varying confounder and/or a time-invariant confounder.

Two of the above studies recommend multiple imputation as the best way of handling missing data; however, there are some important limitations. Each simulation only considered missingness in a confounder or in the treatment; most of the studies only considered up to 50% missingness; and none of the simulations included a time-varying outcome. For these reasons, it is unclear how the different methods studied above would perform when applied to the data in this project.
CHAPTER III

APPLICATION OF A MARGINAL STRUCTURAL MODEL

Study Background

The data for this project are drawn from a multi-year study of asthmatic students attending the Kunsberg School at National Jewish Health. Data were collected on students’ medication use, lung function, asthma symptoms, and environmental exposures. This project focuses on the data from the fifth year of the study, a period of 199 days from October 15th, 2003 to April 30th, 2004. During this school year the study subjects were 43 children with physician diagnosed asthma ranging in age from six to fourteen years. Sixty percent of the subjects were male and 70% were African American. The data structure is one commonly seen in observational studies: there are multiple measurements taken over time, the data collection is intermittent, and many measurements are missing.

<table>
<thead>
<tr>
<th>Table III.1: Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>sex = male (%)</td>
</tr>
<tr>
<td>race = black (%)</td>
</tr>
<tr>
<td>age in years (mean (SD))</td>
</tr>
<tr>
<td>height in cm (mean (SD))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table III.2: Uncollected data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
</tr>
<tr>
<td>fev1_am (mean (SD))</td>
</tr>
<tr>
<td>b asthma1 (mean (SD))</td>
</tr>
<tr>
<td>bdoser1 (mean (SD))</td>
</tr>
</tbody>
</table>

Question of Interest

This project proposes to answer the question: “Does rescue medication (albuterol) use lead to increased lung function (FEV1) in the Kunsberg School asthmatics after accounting for asthma symptoms?”. The variables of interest include a time-varying treatment (rescue medication use); a
time-varying outcome (lung function); baseline measures for subjects’ sex, race, age, and height; and a time-varying confounder (asthma symptoms). Data for the medication use variable were collected each school day (Monday through Friday). Medication use was recorded as the number of times each day a subject used his or her rescue inhaler. Each inhaler included an electronic doser that counted each use. The inhalers were used primarily on an as-needed basis and as such are considered a reflection of a subject’s asthma symptoms. Doser counts were binarized into a yes/no indicator for any medication use for each subject on each day. Data for the asthma variable were collected intermittently in blocks of eight school days in a row. Asthma symptoms were self-reported on a five-point scale that was later binarized. An entry of zero indicates “I did not cough or wheeze last night” or “I coughed and wheezed a little but slept well”. An entry of one indicates “I coughed and wheezed and it woke me at least once in the night or it woke me early”. We hypothesize that the asthma symptoms variable is a time-varying covariate that is both affected by prior medication use and affects future medication use, and over time is both a confounder and a mediator with respect to the relationship between medication use and lung function.

Previous literature suggests there is an estimated 5-10 percent improvement in pulmonary function of asthmatic children after rescue medication use. Mortimer et. al. applied a marginal structural model to data from the Fresno Asthmatic Children’s Environment Study. The authors estimated that rescue medication use was causally related to a 7% increase in pulmonary function when medication use and function were measured on the same day.

The Models

The models for the stabilized weights are logistic models with doser as the outcome:

\[
\text{logit} \ P(doser_{t-1} = 1 | V, \overline{doser}_t) = \alpha_0 + \alpha_1 \text{age} + \alpha_2 \text{height} + \alpha_3 \text{race} + \alpha_4 \text{sex} + \alpha_5 \text{date} + \alpha_6 \text{friday} + \alpha_7 \overline{doser}_t,
\]
\[
\text{logit} \ P(\text{doser}_{t-1} = 1 | V, \overline{\text{doser}}_{t-2}) = \alpha_0 + \alpha_1\text{age} + \alpha_2\text{height} + \alpha_3\text{race} + \alpha_4\text{sex} + \alpha_5\text{date}_{t-1} + \alpha_6\text{friday}_{t-1} + \\
\alpha_7\overline{\text{doser}}_{t-2} + \alpha_8\text{asthma}_{t-1}
\]

The final weighted model is a linear mixed model of the form:

\[
\text{FEV1}_t = \alpha_0 + \alpha_1\text{age} + \alpha_2\text{height} + \alpha_3\text{race} + \alpha_4\text{sex} + \alpha_5\text{date}_{t-1} + \alpha_6\text{friday}_{t-1} + \alpha_7\text{doser}_{t-1} + \alpha_8\overline{\text{doser}}_{t-2}
\]

with a variance components structure for the R matrix. Empirical standard errors that are robust to misspecification of the correlation structure were used to generate p-values and 95% confidence intervals.

The outcome measure, FEV1, was collected in the mornings. Because we wanted to ensure that the treatment, doser, preceded the outcome, a variable for doser lagged one day was used as a predictor instead of the same day’s doser measurement. Similarly, to ensure the value for the time-dependent confounder, asthma symptoms, preceded both the treatment and the outcome, a variable for asthma symptoms lagged one day was used. This approach guaranteed that the three time-varying covariates occurred in an order that made sense causally. However, allowing the treatment and outcome to occur one day apart may result in a weaker estimate of the effect of interest that we would have seen if they were measured on the same day.

The models for this project incorporate treatment history in the form of several additional lagged treatment variables. The number of previous treatment days was determined via a series of logistic regression models with treatment as the outcome. Lagged treatment variables were added to the model sequentially until they were no longer statistically significant predictors of treatment. Doser measurements lagged two days and three days were both significant predictors of treatment; doser lagged four days was not. Therefore, variables for doser lagged two days and three days were included in the weights models and the outcome model.

In a 2008 master’s thesis by Lu\textsuperscript{5}, the author discusses an application of a marginal structural model to the Kunsberg data. The gaps in the weights created by the intermittent data were filled using a
last value carried forward (LVCF) approach: during the compounding of the weights, each time an individual probability was missing, the most recent value for the weight was carried forward and compounding continued. Using this approach, Lu shows that a repeated measures MSM produces estimates that differ from those produced by a standard linear mixed model that adjusts for the time-varying confounder. This project uses the LVCF approach developed by Lu as a starting point, and compares different methods of managing the missingness in the weights caused by intermittent collection in this data.

**Missing Data Methods**

**Imputation of Individual Probabilities and Weights**

The Kunsberg data were intentionally collected intermittently, so most of the ‘missing’ values are not missing in the strict sense. Because of this, it is reasonable to assume the missingness in the data is ignorable. In this case, the best approach may be to use the available case data and fill in any gaps in the weights. These methods are ways of filling in the gaps without requiring any imputation of the data. They present alternatives to the LVCF approach presented in Lu.

Available case data was fit within the MSM with IPTW framework to create stabilized probability of treatment weights as described above. Missing weights caused by missingness in the data were filled in using various methods. The final model was a weighted generalized linear mixed model (GLMM) with FEV1 as the outcome, medication use as the treatment, treatment history, and baseline covariates for race, sex, age, and height. Methods used to fill in the weights included last value carried forward (LVCF) and restarting the weights at one each time there is a gap. Alternatively, gaps in the individual probabilities were filled using (1) average value by subject, (2) average of the two probabilities on either side of the gap, (3) linear interpolation, and (4) the average of randomly generated values. These filled individual probabilities were then used to create the weights. Each method used to fill in the IP produced extreme weight values, and weights greater than 100 were truncated.
Multiple Imputation of Data

Multiple imputation (MI) is a method used for filling in missing data that is widely used in longitudinal studies. It consists of three general steps: First, generate multiple copies of a dataset with missing values imputed using an appropriate model incorporating random variation. Second, analyze each completed dataset separately using a model appropriate to the question of interest. Finally, combine the resulting estimates and standard errors according to Rubin’s rules. There are two general approaches to multiple imputation: joint modelling (JM) and fully conditional specification (FCS). FCS is a more flexible approach that uses univariate conditional distributions for each incomplete variable, allowing for a different imputation model to be selected for each variable. This is the approach that was used for this project.

Imputations were performed in SAS v. 9.4 using PROC MI with an FCS statement for each variable to be imputed. This procedure imputes variables with missing data sequentially in an order specified by the user. First the binary asthma symptoms variable was imputed using a logistic regression model:

$$\text{logit } P(\text{asthma} = 1 | \text{covariates}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{height} + \beta_3 \text{race} + \beta_4 \text{sex}. $$

Then the doser (medication use) variable was imputed using a logistic regression model that included the imputed asthma variable:

$$\text{logit } P(\text{doser} = 1 | \text{covariates}) = \beta'_0 + \beta'_1 \text{age} + \beta'_2 \text{height} + \beta'_3 \text{race} + \beta'_4 \text{sex} + \beta'_5 \text{asthma}. $$

Finally the FEV1 (lung function) variable was imputed using a linear regression model that incorporated both the imputed asthma and doser variables:

$$E[\text{FEV1} | \text{covariates}] = \beta''_0 + \beta''_1 \text{age} + \beta''_2 \text{height} + \beta''_3 \text{race} + \beta''_4 \text{sex} + \beta''_5 \text{asthma} + \beta''_6 \text{doser}. $$

This order was chosen based on the hypothesized causal pathways for the three variables.

When applying multiple imputation methods in the longitudinal setting, repeated measurements of time-dependent variables are considered distinct variables and the dataset is often
reshaped to wide format. However, when the number of repeated measurements is large, MI may encounter non-convergence due to over-fitting or multicollinearity. This was true for the Kunsberg data, so MI on the wide format dataset could not be completed. One suggested alternative approach is to perform the MI with the dataset in long format and include an indicator variable for the cluster in the MI models. For this project, three variations on the imputation models were compared: (1) imputation models as written above, (2) the above models with an indicator for cluster included in each, and (3) the above models with average values of each variable by cluster included in each. Five imputations were run for each scenario.

Results

Imputation of Individual Probabilities and Weights

The methods for filling in gaps in the weights without imputing the data produced varied results. Filling in the individual probabilities with the average IP by subject produced a statistically significant estimate for the relationship between doser and FEV1 ($\beta = 0.099$, $p = 0.025$). This estimate indicates an increase in FEV1 of about 6% on days after subjects used their medication. The result from filling in the IP with randomly generated values was also statistically significant ($\beta = 0.240$, $p = 0.024$). The estimate indicates an increase in FEV1 of about 15% on days after subjects used their medication. Estimates from the other methods ranged from -0.004 to 0.197 and were not statistically significant. See Table III.3 for estimates, 95% confidence intervals, and approximate percent improvement in FEV1 after treatment for each method. The estimates for percent improvement are based on the average FEV1 across all subjects and time points.

Multiple Imputation of Data

Results for all three MI scenarios were similar. The final estimate for the effect of doser on FEV1 was consistently small and consistently statistically insignificant. Estimates for the effect of asthma symptoms on doser in the weights model were also smaller than the estimate from the available case
data. Under imputation scenario (2) the weights model estimated a negative effect of asthma symptoms on doser, but the effect was not statistically significant. Table III.3 shows pooled estimates for the effect of doser on FEV1 and ranges of estimates for asthma symptoms on doser across the imputations. The results suggest there is no significant effect of doser on FEV1, but it is unclear whether the multiple imputation method performed is optimal for the structure of the data.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>95% CI</th>
<th>Approximate improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVCF</td>
<td>0.197</td>
<td>(-0.019, 0.413)</td>
<td>11.9</td>
</tr>
<tr>
<td>restart</td>
<td>0.076</td>
<td>(-0.020, 0.171)</td>
<td>4.6</td>
</tr>
<tr>
<td>IP fill (1)</td>
<td>0.099</td>
<td>(0.012, 0.186)</td>
<td>6.0</td>
</tr>
<tr>
<td>IP fill (2)</td>
<td>-0.004</td>
<td>(-0.091, 0.083)</td>
<td>-0.2</td>
</tr>
<tr>
<td>IP fill (3)</td>
<td>-0.004</td>
<td>(-0.093, 0.086)</td>
<td>-0.2</td>
</tr>
<tr>
<td>IP fill (4)</td>
<td>0.24</td>
<td>(0.031, 0.449)</td>
<td>14.5</td>
</tr>
<tr>
<td>MI Scenario (1)</td>
<td>0.007</td>
<td>(-0.028, 0.043)</td>
<td>0.4</td>
</tr>
<tr>
<td>MI Scenario (2)</td>
<td>0.014</td>
<td>(-0.036, 0.064)</td>
<td>0.8</td>
</tr>
<tr>
<td>MI Scenario (3)</td>
<td>-0.003</td>
<td>(-0.047, 0.042)</td>
<td>-0.2</td>
</tr>
<tr>
<td>Naïve</td>
<td>-0.029</td>
<td>(-0.092, 0.076)</td>
<td>-1.8</td>
</tr>
</tbody>
</table>
Figure III.1: Estimates and 95% confidence intervals for effect of interest
CHAPTER IV

SIMULATION

The simulation was conducted using the free statistical software R version 4.0.2 (https://www.R-project.org/). Nine cases were tested, each one a unique combination of values for n, T, and percent missingness. Number of subjects (i = 1,⋯,n) was either 50 or 500. Number of time points (t = 1,⋯,T) was either 20 or 200. Percent missingness for L, A, and Y, respectively, was 80/30/60 (closely mimicking the Kunsberg data), 40/15/30, or 10/10/10. One thousand datasets were simulated for each case.

Covariates for the outcome Y included a binary variable for treatment (1=treated, 0=not treated), a binary variable for the previous day’s treatment, and a binary time-varying confounder L. Each time-varying confounder value, L_{t,i}, was simulated as follows:

\[ L_{t,i} \sim \text{Bernoulli}(p_{t,i}), \]

where the link function is a logit:

\[ p_{t,i} = P(L_{t,i} = 1 | \beta x_{t,i}) = \frac{\exp (\beta x_{t,i})}{1 + \exp (\beta x_{t,i})}, \]

and

\[ \beta x_{t,i} = \beta_0 + \beta_1 A_{t-1,i} + \beta_2 L_{t-1,i} + \beta_3 Y_{t-1,i}. \]

Each treatment value, A_{t-1,i}, was simulated as follows:

\[ A_{t,i} \sim \text{Bernoulli}(p_{t,i}), \]

where the link function is a logit:

\[ p_{t,i} = P(A_{t,i} = 1 | \alpha x_{t,i}) = \frac{\exp (\alpha x_{t,i})}{1 + \exp (\alpha x_{t,i})}, \]

and

\[ \alpha x_{t,i} = \alpha_0 + \alpha_1 A_{t-1,i} + \alpha_2 L_{t,i}. \]

Continuous outcome values, Y_{ij}, were simulated as:
\[
Y_{t,i} = \gamma_0 + \gamma_1 A_{t,i} + \gamma_2 A_{t-1,i} + \gamma_3 L_{t,i} + \varepsilon_{t,i} + b_{0i}
\]

where:
\[
\varepsilon_{t,i} \sim N(0, \sigma^2), \ b_{0i} \sim N(0, \sigma^2).
\]

Figure IV.1 shows the directed acyclic graph for the simulated data.

![Directed acyclic graph](image)

Figure IV.1: Directed acyclic graph for the relationships between L, A, and Y

The true coefficient values for these variables were set as \( \beta = (0, -0.5, 0.1, -0.3) \), \( \alpha = (0, 1, 1) \), and \( \gamma = (2, 0.5, -0.75, 0.1) \). Random intercept values for each subject were drawn from a normal distribution with mean 0 and variance 1. See Table IV.1 for a summary of the parameter values for all 9 cases.

Variables for L, A, and Y with some data points missing were created in the dataset by randomly deleting data points according to the three missing data schemes described above. Six models were then fit to each dataset: two on the L, A, and Y variables with no missing data points (the ‘full’ data), and four on the variables with missing data points. The full data were fit using a MSM with IPTW and with a linear mixed model adjusting for L as a confounder (the ‘naïve’ model). Four marginal structural models were fit on the variables with missing data points: one filled in gaps in the weights by carrying the last weight forward, one filled in gaps in the weights by restarting the weights at one each time there was a gap, one filled in the individual probabilities randomly and then used the filled IP to create weights, and one filled in the IP with the average IP by subject and then used the filled IP to create weights. Any weights
greater than 100 were truncated to avoid convergence issues. Empirical standard errors were used to create confidence intervals and p-values for each model.

<table>
<thead>
<tr>
<th>Case</th>
<th>No. subjects (n)</th>
<th>No. time points (t)</th>
<th>% missing L</th>
<th>% missing A</th>
<th>% missing Y</th>
<th>true value $\beta_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>20</td>
<td>80</td>
<td>30</td>
<td>60</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>20</td>
<td>40</td>
<td>15</td>
<td>30</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>200</td>
<td>80</td>
<td>15</td>
<td>30</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>200</td>
<td>40</td>
<td>30</td>
<td>60</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>200</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>20</td>
<td>80</td>
<td>30</td>
<td>60</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>500</td>
<td>20</td>
<td>40</td>
<td>15</td>
<td>30</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>500</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Beta coefficients for the effect of A on Y and 95% confidence intervals (CI) for each model were saved and summarized. Percent coverage for each estimated beta coefficient was calculated as the proportion of generated 95% CI that included the true beta. Bias in the coefficient estimates was calculated by subtracting the true beta value from the estimated beta value for each generated value of each coefficient. Type II error was calculated as the proportion of 95% CI that covered 0. In addition to the main simulation results, one dataset from each case was selected randomly for multiple imputation. Imputation was performed on each dataset using the mean values of each variable by subject as described above.

**Results for Main Simulation**

Mean values of the estimated $\beta_A$, mean values for bias, mean values for MSE, percent coverage, and percent type II error are summarized in Table IV.2. Selected boxplots show the variation in the estimated beta coefficients for cases 1, 4, and 7 (see Appendix).

As expected, mean bias and type II error tend to decrease and coverage tends to increase for all four models on the missing data as the percent of missing data decreases. Type II error is an issue for all
methods when the numbers of subjects and time points are smaller and the percent of missing data is the highest.

Across most of the cases, the method of filling in the individual probabilities randomly tends to have the highest mean bias and the highest percent type II error. When the number of time points is 20, the random IP method also has the lowest coverage and the highest mean squared error (MSE), though when the number of time points is 200 the last value carried forward (LVCF) method has low coverage and high MSE compared to the other methods. Interestingly, coverage is noticeably lower and MSE higher in the model on the full data when the number of time points is 200 compared to when the number of time points is 20. The MSM outperforms the naïve model in the cases with 20 time points but has greater mean bias and lower percent coverage when the number of time points is 200.

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.539 (0.037)</td>
<td>0.525 (0.048)</td>
<td>0.591 (0.233)</td>
<td>0.546 (0.185)</td>
<td>0.413 (0.307)</td>
<td>0.532 (0.302)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.039</td>
<td>0.025</td>
<td>0.091</td>
<td>0.046</td>
<td>-0.087</td>
<td>0.032</td>
</tr>
<tr>
<td>MSE</td>
<td>0.003</td>
<td>0.003</td>
<td>0.063</td>
<td>0.036</td>
<td>0.102</td>
<td>0.092</td>
</tr>
<tr>
<td>% coverage</td>
<td>84.9</td>
<td>97.5</td>
<td>89.9</td>
<td>93.7</td>
<td>72.8</td>
<td>80.2</td>
</tr>
<tr>
<td>% type II error</td>
<td>0.0</td>
<td>0.0</td>
<td>21.1</td>
<td>13.6</td>
<td>38.8</td>
<td>24.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 2</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.521 (0.037)</td>
<td>0.503 (0.047)</td>
<td>0.521 (0.103)</td>
<td>0.511 (0.100)</td>
<td>0.446 (0.177)</td>
<td>0.509 (0.124)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.021</td>
<td>0.003</td>
<td>0.021</td>
<td>0.011</td>
<td>-0.054</td>
<td>0.009</td>
</tr>
<tr>
<td>MSE</td>
<td>0.002</td>
<td>0.002</td>
<td>0.011</td>
<td>0.01</td>
<td>0.034</td>
<td>0.016</td>
</tr>
<tr>
<td>% coverage</td>
<td>93.9</td>
<td>97.1</td>
<td>95.5</td>
<td>95.6</td>
<td>83.6</td>
<td>92.1</td>
</tr>
<tr>
<td>% type II error</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>17.2</td>
<td>1.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 3</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.557 (0.037)</td>
<td>0.537 (0.050)</td>
<td>0.547 (0.061)</td>
<td>0.548 (0.064)</td>
<td>0.508 (0.135)</td>
<td>0.544 (0.069)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.057</td>
<td>0.037</td>
<td>0.047</td>
<td>0.048</td>
<td>0.008</td>
<td>0.044</td>
</tr>
<tr>
<td>MSE</td>
<td>0.005</td>
<td>0.004</td>
<td>0.006</td>
<td>0.006</td>
<td>0.018</td>
<td>0.007</td>
</tr>
<tr>
<td>% coverage</td>
<td>71.6</td>
<td>94.4</td>
<td>93.1</td>
<td>93.6</td>
<td>88.7</td>
<td>93.4</td>
</tr>
<tr>
<td>% type II error</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>4.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Case 4</td>
<td>Naïve model</td>
<td>MSM (full)</td>
<td>MSM (LVCF)</td>
<td>MSM (restart)</td>
<td>MSM (random IP)</td>
<td>MSM (avg IP)</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
<td>---------------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.504 (0.011)</td>
<td>0.491 (0.037)</td>
<td>0.568 (0.253)</td>
<td>0.441 (0.049)</td>
<td>0.377 (0.087)</td>
<td>0.395 (0.124)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.004</td>
<td>-0.009</td>
<td>0.068</td>
<td>-0.059</td>
<td>-0.122</td>
<td>-0.105</td>
</tr>
<tr>
<td>MSE</td>
<td>0</td>
<td>0.001</td>
<td>0.069</td>
<td>0.006</td>
<td>0.023</td>
<td>0.026</td>
</tr>
<tr>
<td>% coverage</td>
<td>94.1</td>
<td>89.7</td>
<td>25.4</td>
<td>77.7</td>
<td>61</td>
<td>67.3</td>
</tr>
<tr>
<td>% type II error</td>
<td>0</td>
<td>0</td>
<td>3.8</td>
<td>0</td>
<td>1.4</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 5</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.500 (0.011)</td>
<td>0.487 (0.040)</td>
<td>0.491 (0.096)</td>
<td>0.479 (0.031)</td>
<td>0.391 (0.060)</td>
<td>0.446 (0.100)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0</td>
<td>-0.013</td>
<td>-0.009</td>
<td>-0.02</td>
<td>-0.109</td>
<td>-0.054</td>
</tr>
<tr>
<td>MSE</td>
<td>0</td>
<td>0.002</td>
<td>0.009</td>
<td>0.001</td>
<td>0.015</td>
<td>0.013</td>
</tr>
<tr>
<td>% coverage</td>
<td>95.8</td>
<td>86.5</td>
<td>86.1</td>
<td>91.5</td>
<td>44.9</td>
<td>63.2</td>
</tr>
<tr>
<td>% type II error</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 6</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.502 (0.012)</td>
<td>0.491 (0.038)</td>
<td>0.494 (0.044)</td>
<td>0.497 (0.022)</td>
<td>0.419 (0.048)</td>
<td>0.489 (0.057)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.002</td>
<td>-0.009</td>
<td>-0.006</td>
<td>-0.003</td>
<td>-0.081</td>
<td>-0.01</td>
</tr>
<tr>
<td>MSE</td>
<td>0</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>% coverage</td>
<td>94.1</td>
<td>88.2</td>
<td>93.3</td>
<td>95.2</td>
<td>54.1</td>
<td>78.2</td>
</tr>
<tr>
<td>% type II error</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 7</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.522 (0.012)</td>
<td>0.506 (0.016)</td>
<td>0.513 (0.069)</td>
<td>0.486 (0.054)</td>
<td>0.401 (0.125)</td>
<td>0.506 (0.156)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.022</td>
<td>0.006</td>
<td>0.013</td>
<td>-0.014</td>
<td>-0.099</td>
<td>0.006</td>
</tr>
<tr>
<td>MSE</td>
<td>0.001</td>
<td>0</td>
<td>0.005</td>
<td>0.003</td>
<td>0.025</td>
<td>0.024</td>
</tr>
<tr>
<td>% coverage</td>
<td>59.2</td>
<td>97.4</td>
<td>95.8</td>
<td>95.3</td>
<td>80.7</td>
<td>89.3</td>
</tr>
<tr>
<td>% type II error</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.4</td>
<td>6.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 8</th>
<th>Naïve model</th>
<th>MSM (full)</th>
<th>MSM (LVCF)</th>
<th>MSM (restart)</th>
<th>MSM (random IP)</th>
<th>MSM (avg IP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate (mean (sd))</td>
<td>0.528 (0.012)</td>
<td>0.511 (0.016)</td>
<td>0.528 (0.033)</td>
<td>0.517 (0.032)</td>
<td>0.445 (0.071)</td>
<td>0.514 (0.049)</td>
</tr>
<tr>
<td>Bias (mean)</td>
<td>0.028</td>
<td>0.011</td>
<td>0.028</td>
<td>0.017</td>
<td>-0.055</td>
<td>0.014</td>
</tr>
<tr>
<td>MSE</td>
<td>0.001</td>
<td>0</td>
<td>0.002</td>
<td>0.001</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>% coverage</td>
<td>40</td>
<td>94.7</td>
<td>92.6</td>
<td>94.5</td>
<td>83.3</td>
<td>95.3</td>
</tr>
<tr>
<td>% type II error</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure IV.2: Comparison of percent coverage and MSE for simulation results. Plots in the top row show results from cases with the highest percentage of missing data (cases 1, 4, and 7); plots in the middle row show results from cases with the mid-range percentage of missing data (cases 2, 5, and 8); and plots in the bottom row show results from cases with the lowest percentage of missing data (cases 3, 6, and 9). Different combinations of number of subjects and time points are shown on the x-axis.

Results for Multiple Imputation on Imputed Data

The results for multiple imputation on the simulated data showed a consistent pattern across all cases. Analysis of the imputed variables gave estimates for $\beta_A$ that were lower than the estimates for the
full data; and the higher the percentage of missingness, the lower the estimate. See results for all nine cases in Table IV.3. Unlike the MI performed on the Kunsberg data, the reduced estimates from the simulated data remained statistically significant; however, this could be due to a stronger effect of A on Y in the simulated data.

<table>
<thead>
<tr>
<th>Case</th>
<th>( \beta ) full data</th>
<th>95% CI full data</th>
<th>( \beta ) imputed data</th>
<th>95% CI imputed data</th>
<th>% decrease in estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.525</td>
<td>(0.401, 0.650)</td>
<td>0.285</td>
<td>(0.177, 0.392)</td>
<td>45.7</td>
</tr>
<tr>
<td>2</td>
<td>0.591</td>
<td>(0.485, 0.697)</td>
<td>0.374</td>
<td>(0.248, 0.500)</td>
<td>36.7</td>
</tr>
<tr>
<td>3</td>
<td>0.539</td>
<td>(0.447, 0.632)</td>
<td>0.454</td>
<td>(0.352, 0.556)</td>
<td>15.8</td>
</tr>
<tr>
<td>4</td>
<td>0.506</td>
<td>(0.423, 0.589)</td>
<td>0.282</td>
<td>(0.197, 0.366)</td>
<td>44.3</td>
</tr>
<tr>
<td>5</td>
<td>0.445</td>
<td>(0.380, 0.510)</td>
<td>0.348</td>
<td>(0.245, 0.451)</td>
<td>21.8</td>
</tr>
<tr>
<td>6</td>
<td>0.491</td>
<td>(0.447, 0.535)</td>
<td>0.440</td>
<td>(0.338, 0.542)</td>
<td>10.4</td>
</tr>
<tr>
<td>7</td>
<td>0.522</td>
<td>(0.489, 0.556)</td>
<td>0.232</td>
<td>(0.182, 0.283)</td>
<td>55.6</td>
</tr>
<tr>
<td>8</td>
<td>0.504</td>
<td>(0.470, 0.538)</td>
<td>0.365</td>
<td>(0.314, 0.415)</td>
<td>27.6</td>
</tr>
<tr>
<td>9</td>
<td>0.532</td>
<td>(0.490, 0.573)</td>
<td>0.420</td>
<td>(0.376, 0.464)</td>
<td>21.1</td>
</tr>
</tbody>
</table>
CHAPTER V
DISCUSSION

The simulation results suggest that filling in the individual probabilities randomly before creating the weights is not a good choice for the Kunsberg data due to the variability and bias in the estimates. Neither is the multiple imputation method, because of the dramatic decrease in the estimate of the effect of interest when the percentage of missing data was high. In the cases with a small number of subjects and large number of time points, the LVCF method had poor coverage and more variability in the estimates than the random IP method. These results suggest the method of restarting the weights is the most adequate of those explored here, as it consistently showed the smallest mean bias and highest percent coverage of the models fit on the variables with missing data points. The estimate of the effect of doser on FEV1 produced by restarting the weights was not statistically significant ($\beta = 0.076, p = 0.122$). However, this estimate suggests an increase in FEV1 of about 4.6% on days after subjects used their medication, which is a reasonable estimate given results from previous literature\textsuperscript{12}.

Limitations

The above conclusions rest on the assumption that the simulated data reflect the structure of the Kunsberg data. We have assumed statistically significant effects of asthma symptoms on medication use, medication use on asthma symptoms, asthma symptoms on FEV1, and medication use on FEV1. We have also assumed the missingness is ignorable. These assumptions are unverified, and if false may invalidate the simulation results.

The MSM with IPTW approach has several notable advantages and disadvantages. It is easy to understand and to implement, and only the final MSM and the models for the weights need to be specified. However, the weighting can be unstable and inefficient when there are extreme weights\textsuperscript{15}. In cases where extreme weights are present, weight truncation may improve estimation of the marginal
treatment effect by increasing coverage and decreasing bias\textsuperscript{11}, but there is currently no set standard for weight truncation known to produce optimal results.

**Future Directions**

Several of the methods for filling in missing weights employed in this project produced extreme weight values. In the future it may be useful to experiment with different weight truncation or normalization methods to see how the model results are affected. It may also be useful to introduce non-ignorable missingness and add an unmeasured confounder to the simulated data, given that our assumptions of ignorable missingness and no unmeasured confounding are untestable. Experimenting with a wider variety of values for number of subjects and number of time points could help to narrow down more specific instances when the naïve model outperforms the MSM.
REFERENCES


FIGURE A.1: Distribution of estimates for case 1
Figure A.2: Distribution of estimates for case 4

Figure A.3: Distribution of estimates for case 7