EVALUATION OF DYNAMIC CLUSTER-BASED IMPUTATION FOR NON-IGNORABLE MISSING DATA IN CLUSTERED RANDOMIZED DATA

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE

by
Christopher Dzura
December 2014

Committee Members:

Dr. Adam Davey, Thesis Advisor _____
Department of Public Health, Temple University

Dr. Deborah Nelson _____
Department of Public Health, Temple University

Dr. Mark Weir _____
Department of Public Health, Temple University

ABSTRACT

There is growing recognition that ignoring missing data requires stronger assumptions than addressing them. Numerous effective methods have emerged for handling missing data under so-called ignorable missing data conditions, effective treatment of non-ignorable missing data (missing not at random, MNAR) remains an open problem, typically requiring that missing data be modeled explicitly. Most previous research applies a single statistical model to the entire data set, although some previous research suggests that imputations based on a subset of similar cases may prove more effective with MNAR data. The primary aim of this study was to determine if imputing missing data by a method which utilizes only local information could provide results comparable to methods based on explicit modeling of missing data. This thesis reports the results of 4 experiments evaluating dynamic cluster-based imputation (DCI) to impute missing post-test data in simulated cluster-randomized trial data under conditions that previous research showed no standard methods to be effective. This method identifies a set of statistical proximal observations (indexed by K) and uses a subset of them (indexed by R) to perform imputations. Both parameters can be tuned to a specific problem with previous research suggesting that K=50 and R=9 was optimal for most applications. Compared to common methods for handling non-ignorable missing data (listwise deletion, LOCF, pattern-mixture modeling, Diggle-Kenward modeling), results of the experiments suggest that indicators variables for the pattern of missing observations was included in the imputation model, DCI had comparable bias and coverage rates to “best practice” methods of Diggle-Kenward and pattern mixture models, but at the cost of larger

standard errors. Thus results support for the idea that imputation based on a subset of similar observations could be more accurate than imputation using all cases.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

Missing data are an unavoidable part of conducting public health research. The reasons data can be missing are diverse and vary depending on the aims and design of the study. Survey respondents can choose not to answer questions or ignore the entire questionnaire, known as item nonresponse and unit nonresponse, respectively. In longitudinal studies, missing data are broken down into two categories, monotone and nonmonotone. Monotone (called dropout for the rest of this work) missing data occur when data collection ceases earlier than expected; subjects withdrawing their consent or dying over the course of the study are examples of monotone attrition. Nonmonotone (called intermittent missing for the rest of this work) categorizes unpredictable or sporadically missing observations. Public health studies and clinical trials are frequently longitudinal with repeated measurements and commonly encounter both monotone and nonmonotone patterns of missing data.

An early attempt in social sciences to address missing data involved coding missing observations with dummy variables (Cohen and Cohen, 1983). Although considered a crude method by contemporary standards, this was recognition that missing data cannot simply be discarded or ignored. Over time, computers have increased the ease in using more sophisticated techniques to treat incomplete records. As a result of the proliferation in statistical software and computing power, there is a greater expectation that missing data will be handled with an adequate technique by researchers.

Answering the question of why it is necessary to address missing data before conducting analyses is straightforward. Missing data represent unknown attributes from

the sample researchers are working with. These omissions are essentially placeholders for valuable data that were not collected; less is known about the sample than under ideal conditions. The goal of many public health studies is to make inferences about a greater population based on a sample, this becomes more difficult when information from the sample is incomplete. Any treatment applied to a data set with missing data can have an impact on the results; even choosing to ignore observations with missing data can have an effect. When the inferences drawn from a sample are not truly representative of the population they belong to, bias has been introduced. The adverse influence bias can have on inferences is a strong motivation to correctly handle missing data.

When it comes to handling missing data, identifying an appropriate technique involves diagnosing the type of missing data to be analyzed. The types of missing data and common methods for handling them are explained in Chapter 2: Background. The goal of this thesis is to determine, via a series of research questions, if a technique developed from computer science can be used to accurately impute missing values in a large data set. If successful, this new imputation technique would give researchers an additional method for treating missing data. Details of the imputation technique, the data set utilized and research questions are described in more detail in Chapter 3: Methods. The experimental results are documented in Chapter 4: Results and examined in Chapter 5: Discussion. Chapter 6 is the conclusion which summarizes the work.

CHAPTER 2: BACKGROUND

Missing data are an unavoidable part of conducting public health research. The causes of missing data are diverse and reflect the type of research being conducted. Careful design of data collection techniques can reduce, but not eliminate, the frequency of missing data (National Research Council, 2010). Suggestions include utilizing designs that minimize dropout and continuing to collect data on participants who choose to discontinue treatment. In his contribution to *Longitudinal Data Analysis*, Little (2008) stresses the importance of limiting design flaws that invite missing data, "since any method for compensating for missing data requires unverifiable assumptions that may or may not be justified. Since data are still likely to be missing despite these efforts, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible, since this information helps to model the missing data mechanism when the incomplete data are analyzed." (p. 409) A description of each missing data mechanism is provided in the following section.

Types of Missing Data

Little and Rubin (2002) classified three mechanisms known as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). These three taxonomic labels provide researchers a common terminology when discussing missing values based on their relationship to observed and unobserved variables.

MCAR: missing completely at random

When the probability that an observation is missing Pr(r) does not depend on either the observed (Y_{obs}) or unobserved (Y_{miss}) variables data are said to be MCAR.

$Pr(r \mid Y_{obs}, Y_{miss}) = Pr(r)$

Data are classified as missing completely at random (MCAR) when their missingness is not explained by the observed data or the missing data; this pattern is truly unpredictable. Missing cases are similar to non-missing cases in terms of their parameters and distribution. Highly unlikely in many real-life applications, this situation is considered the easiest to accommodate when dealing with missing data.

MAR: missing at random

When the probability that an observation is missing Pr(r) depends only on the values of observed variables data are said to be MAR.

$Pr(r \mid Y_{obs}, Y_{miss}) = Pr(r \mid Y_{obs})$

Data are classified as missing at random (MAR) when their missingness is explained only by the observed data. Missing cases are not similar to non-missing cases but factor(s) which explain their difference(s) are observable in the collected data. Missing at random is a more realistic situation than missing completely at random since there are likely some differences between the missing and non-missing cases.

MNAR: missing not at random.

When the probability that an observation is missing Pr(r) depends on the values of unobserved variables data are said to be MNAR.

$Pr(r \mid Y_{obs}, Y_{miss}) = Pr(r \mid Y_{miss})$

Data are classified as missing not at random (MNAR) when their missingness is explained by the unobserved data. Missing cases are not similar to non-missing cases and there is a relationship between a missing value and the likelihood it is missing. MNAR missingness is the most challenging to treat because the underlying mechanism, by its nature, cannot be observed and any assumptions made about modeling the missing data are untestable. In practice, it is impossible to distinguish between MAR and MNAR data.

Mallinckrodt and colleagues (2003) illustrated how these three different patterns of missing data can occur during the same clinical trial:

> "MCAR data may arise from a patient who dropped out because he relocated and was too far away from the investigative site to participate in the trial. Dropout was not in any way related to the outcome of interest. An example of MAR data could be an instance in which a patient was observed to be doing poorly and then the physician and/or the patient decided to discontinue participation. In this case, dropout was related to the outcome of interest, but the observed data explained the dropout. An example of MNAR data could be a patient who had been doing well until midway in a trial and was then lost to follow up because after the last observed visit the patient relapsed into a worsened condition. Again, dropout was related to the outcome of interest, but in this case the observed data did not explain (predict) the dropout, and the unobserved data held information not foreseen by the observed data."

Although the result is the same for all three cases, missing data as a consequence of subject dropout, this example demonstrates how varied the reasons for missing data can be. Frequently, missingness is related to the outcome of interest, meaning data are rarely MCAR. Assuming data are MAR and that misingness can be explained in the observed data is more plausible than MCAR (Little and Rubin, 2002). MNAR data are difficult to predict but randomized trials are a setting where an abundance of measurements are taken; it is then more likely that the reason data are missing can be explained with observed data, reducing the likelihood of MNAR data.

Ignorable and Non-ignorable Mechanisms

When the likelihood of a value being missing can be attributed to random chance (MCAR) or to observable data (MAR), these two scenarios are collectively known as ignorably missing data. In cases of ignorably missing data the mechanism describing the pattern of missingness is at least partially known. Labeling missing data as ignorably missing does not mean those unknown values require no treatment when present; ignorable data require less assumptions for modeling their missing data mechanism. Alternatively, MNAR data are known as non-ignorable because the mechanism explaining their missingness pattern is unknown and unobservable. This data is missing in an unmeasured fashion and must be addressed before any reliable analysis can be conducted. In cases of non-ignorable data where the likelihood of a missing observation depends on its true (and unmeasured) value , sensitivity analyses are able to help reduce the uncertainty of modeling an unobserved phenomenon such as this.

Approaches to Treating Missing Data

Varied in complexity and prerequisite conditions for use, there are many diverse techniques available for handling missing data. Deciding which approach is optimal for a data set depends on many factors, including the suspected type of missing data and the resources available to the researcher. The techniques described below are some of the most commonly used methods of treating missing data.

Complete Case Analyses

Complete case analyses ignore the records or subjects with missing data. These incomplete cases are omitted from any statistical procedures because they lack some

information required to make inferences. There are many situations in which a complete case analysis can result in misleading conclusions so their use is not recommended (Graham, 2009).

Listwise Deletion

A simple solution when missing data are encountered is to remove it. Discarding all cases with missing data is known as listwise deletion. Likely to produce unreliable or biased results, the appeal of deletion is the ease of its implementation. Except in rare instances of MCAR data, removing incomplete cases will bias the results because of the fundamental difference between complete and incomplete records. Even when used on MCAR data, listwise deletion results in a loss of statistical power due to the reduction in effective sample size. Additionally, any partial information provided by incomplete cases will be lost when choosing listwise deletion.

Weighting Methods

A noted disadvantage of typical complete case analyses are the bias of their parameter estimates when data are not MCAR. Weighting methods like Inverse Probability Weighting (IPW) attempt to reweight the complete cases in such a way that it restores the representativeness of the original data set. By predicting the probability of missingness of one variable with the rest of the variables in the data set, the inverse of these predicted probabilities can be used as the case weightings. This method increases in complexity as the number of variables with missing data increases. Weighting data for analytical inference in this way is controversial because it can be viewed as a manipulation of the observed data.

Selection and Pattern-Mixture Models

Longitudinal studies rely on repeated measurements to track the change of subject results over time. These changes can be modeled in terms of fixed effects and random effects (experiment-specific factors and subject-specific factors, respectively). Selection and pattern-mixture models are different factorizations of joint distributions that are used depending on the context of data missingness. These models are used when dropout rates and intermittent missing data would make interpretation of the results difficult; based on how the conditional probabilities are arranged, selection models are appropriate for MAR (missing data are conditioned on observed values) and pattern-mixture for MNAR (data are split into subgroups based on their pattern of missing data) (Yang, Li and Shoptaw, 2008). Analytically, dropout is problematic due to its suspected relationship to the observed and missing values. Diggle and Kenward (1994) proposed a specific type of selection model to describe the association between missing data and other variables in the data set, like treatment received. Since it relates observed values to missing values, this is a model which can be used for handling MNAR dropout.

Selection and pattern-mixture models, two different modeling strategies, can be used in conjunction for conducting a sensitivity analysis. A sensitivity analysis explores the impact of a model and/or selected observations on the inferences made when data are incomplete. (Molenberghs, 2009). The agreement between models with different assumptions and influential observations can provide insight into which models best fit the collected data. Normally untestable, the missingness mechanism of suspected MNAR

data benefits from being evaluated under varied conditions like those found in a sensitivity analysis.

Available Case Analyses

In contrast to complete case analyses that discard incomplete records, available case analyses use all the gathered data when estimating missing values. Incomplete cases are still valuable for the partial information they contain.

Full Information Maximum Likelihood

Full Information Maximum Likelihood (FIML) computes a likelihood function for each case based on its available data. These casewise likelihoods are then aggregated and maximized. This is known as a direct maximization technique because the model parameters are calculated with only the available data; there are other likelihood techniques which include an expectation (imputation) step before parameter maximization. When data are ignorably missing, using the partial information from incomplete cases typically improves parameter estimates. It has been suggested (Arbuckle, 1996) that parameter estimates generated by using FIML can in part depend on the degree of association between observed variables, with greater efficiency in prediction when variables are more strongly correlated. Without an expectation step FIML is less computationally demanding compared to other likelihood techniques, although this method is not commonly available in statistical packages and requires specialized software to run.

Imputation Methods

Unlike techniques from the complete case and available case analysis families described above, imputation techniques generate replacement values for the unobserved data. These techniques have encountered skepticism because it can appear that imputed data are fabricated. Raghunathan (2004) notes two key points which reinforce important limitations of imputed values in a data set:

- Any imputation procedure should not be viewed as a method for recovering the missing values for any given individual.
- Filled-in values for any one subject with missing values should not be considered as microdata for that subject but rather as values that are statistically plausible given other information on that subject.

Imputation is meant to improve the accuracy of the data as a whole and not to predict missing values; individually imputed results are not meant to be perfect substitutes for the value they are replacing.

Single Imputation

Single imputations techniques substitute a missing value with one plausible value. After all missing values have been imputed, standard statistical procedures are performed on the entire data set as if data was fully observed. Single imputations are replicable and deterministic: given the same starting conditions the procedure will provide the same imputed values every time it is performed. Intuitive and commonly available in statistical software, these techniques are frequently used as a primary analysis tool. Although single imputations retain statistical power because sample size remains unchanged, the uncertainty resulting from dealing with missing data is understated. The actual stochastic

realization (Molenberghs et al., 2004) of any missing value depends on unknown mean and variance structures and the effect they have on an imputation operation is minimized by single imputation procedures.

Hot Deck and Cold Deck Imputation

Based on terms that date back to computing with punch cards, hot deck imputations are techniques that use values found within the data set for its imputation. Cold deck imputations use data collected from a different study when imputing missing values. For example, an annual census could use the previous year's data to make inferences about missing respondents. All three single imputation techniques discussed below could be categorized as hot deck since they draw values only from their own observed data.

Last Observation Carried Forward

Last Observation Carried Forward (LOCF) is a single imputation technique used with longitudinal data. A subject's missing values are replaced with the last known (most recently observed) value prior to the missing observation. The rationale behind this method is that the best estimate of a missing value is to substitute the most recently collected value from that subject.

Proponents of this technique state LOCF is a very objective method of data imputation because only observed values are used as substitutes for missing values. There are certain situations, including after many values in a row are imputed with the same value, where the reliability of the method has been called into question. Any potential treatment effect is masked during the period of time imputed values are involved.

Mean Substitution

Mean substitution replaces the missing values of each variable with an average of the non-missing values. This imputation preserves the mean of the sample but the lack of variation in values results in reduced confidence when making inferences. There are numerous disadvantages to using mean substitution including the inability to accommodate categorical variables and a lack of sensitivity to extreme missing observations. Mean substitution also yields biased parameter estimates for all types of missing data, including MCAR; Wothke (2000) remarked that mean substitution results in "very precise estimates of exactly the wrong parameter".

Regression Imputation

Another single imputation technique, regression imputation uses a regression line to predict missing values. With the information available in complete and incomplete cases, each subject has any missing values imputed with a regression model that uses their observed measurements as independent variables coefficients. Fitted values from each regression model are then used to impute the missing values for that variable. This method suffers many of the same drawbacks as mean substitution, namely that there is no variability in the predicted values resulting in over-confidence of the imputation. In contrast to mean substitution, regression imputation makes relationships between variables appear too strong; the lack of any error term results in predicted values that perfectly fit on the regression line every time.

Multiple Imputation

Multiple imputation techniques are based on the same principle as single imputation except that more than one plausible value is considered for each missing value. This set of potential values represents uncertainty about the exact value in question; the final imputed value is an aggregation of all plausible values. Generating varied data sets requires a probability model constructed based on Bayesian inference. After approximating a prior distribution for the unknown parameters, random draws are made on the conditional distribution of the missing data given the observed data.

The advantage of this approach over any single imputation technique is that uncertainty in parameter estimates is accounted for by the inherent variability of creating multiple plausible values. While multiple imputation may sound like a laborious process, high levels of estimation efficiency can be obtained within a limited number of imputations. With *m* imputations and a missing data percentage of (γ), Rubin (1987) showed that parameter estimation efficiency follows the formula:

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

Figure 1 demonstrates the high efficiency that can be achieved for fixed numbers of imputations and rates of missing data. Except in cases with extremely high rates of missing data even very few imputations can result in reliable parameter estimates.

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Figure 1: Parameter Estimation Efficiency for Multiple Imputations
(m)= number of imputations; (γ)=rate of missing data

Multiple Imputation by Chained Equations

Multiple Imputation by Chained Equations (MICE) uses a sequential regression design which iteratively improves the precision of each missing value. After seeding all missing data with a simple imputation like mean substitution, MICE treats each variable like a regression model conditional on all other variables in the data set. This regression model predicts new (and more accurate) values that replace the seeded values. By iteratively rerunning these chained models the regression parameters of each variable eventually converge and stabilize. Regressing each variable in this manner has many practical benefits for imputation: MICE is flexible and can accommodate varied data types with an appropriate regression (e.g. continuous variables with linear regression, binary variables with logistic regression and categorical variables with logit regression) which makes it useful for mixed data that would require special attention with other imputation techniques (Stuart et al., 2009). MICE is only reliably accurate when data are assumed as MAR.

Analysis of Missing Data

Before conducting any analysis with incomplete data, it is important to note that there is not one single best method for handling all instances of missing data. As the three types of missing data have different characteristics and carry different modeling assumptions, it follows that they are best treated with different techniques. The following is a discussion of which techniques work best for each type of missing data.

Missing Completely at Random

When data can be considered MCAR there are many techniques which can handle this type of missingness. This is the only scenario when complete case analyses can return unbiased results because there is no systematic difference between complete and incomplete records. Although complete case techniques can be used, available case techniques leverage more information from the data set and are preferable. In Full Information Maximum Likelihood (FIML) no explicit imputation takes place, hence the amount of information in the data is not overestimated and important model elements, such as mean and variance, are not distorted (Molenberghs et al., 2004). Multiple imputation provides another viable technique for handling MCAR data as well.

Missing at Random

The other type of ignorably missing data, MAR analyses can be handled in a manner similar to MCAR. Complete case techniques are no longer acceptable to use because complete and incomplete records are distinct. FIML is still a valid technique but now has a stricter requirement for implementation: variables which explain nonresponse must be included in the model. The partial information explanatory variables provide is

necessary for accurate likelihood computations. Multiple imputation is an acceptable technique to use for MAR data as well.

Missing Not at Random

Treating MNAR data requires stricter assumptions and fundamentally different techniques than the two former types of missing data. Before handling this type of data, nonresponse must be explicitly modeled because of the untestable relationship between the rate and value of unobserved values. Pattern-mixture models and Diggle-Kenward modeling techniques are frequently used when data are suspected of being nonignorably missing. Correctly modeling the mechanism for missing data is a key part of drawing accurate inferences with MNAR data.

Remarks on Last Observation Carried Forward

Although Last Observation Carried Forward (LOCF) is frequently used to impute data in randomized trials, this technique has drawbacks which warrant discussion. Treatment effect, assessed as change between baseline and study completion, is commonly analyzed as analysis of variance (ANOVA) with missing data imputed by LOCF. LOCF substitutes unmeasured values with the last observed value, resulting in one fixed value imputed for subjects who drop out of a study. It follows that this method is influenced by the timing and reason for dropout; assuming measurements do not change after dropout (or during any unobserved period in the cases of intermittent missingness) is questionable. It seems implausible that missing observations due to study-related factors (MAR or MNAR) would have remained unchanged for the duration of the study, given that subject continued participating. Carrying a fixed value forward

introduces potential bias of treatment effect when the timing and reason for missing values are disregarded.

Despite these methodological drawbacks, LOCF is still commonly selected for treating missing data due to the belief that the results are a conservative estimate of treatment effect. Some of the appeal for this technique also stems from how concrete the imputation appears: the values substituted during LOCF are present in the original data set and not "made up" as the values of other imputations can appear, for example, with mean substitution. The assumption that treatment effect is conservatively estimated when conducting ANOVA LOCF is not based on any theoretical framework and has been documented as increasing the risk of committing type 1 error (Mallinckrodt et al., 2003). Additionally, as a single imputation technique, LOCF has no stochastic component for imputed values and thus overstates the certainty they contribute to the data set. When presented with a clear picture of the bias which can be introduced with LOCF, it is difficult to justify its use, especially when other techniques can be applied that address these shortcomings.

The Impact of Missing Data in Public Health

The aim of a randomized trial is to assess the effectiveness of a treatment, usually compared to a standard approach. Missing data represent scheduled points in time when measurements were not taken and as a consequence makes comparing the effectiveness between treatments more challenging. There are a variety of reasons for missing observations and Khan et al., (2007) demonstrated that a subject's reason for dropout can vary significantly based on disorder (depression, schizophrenia or OCD) and treatment

arm assignment (active drug versus placebo). Dropout motivated by a perceived lack of treatment effect or intolerance to side effects are types of non-ignorably missing data described as "outcome-dependent censoring" by Yang and Shoptaw (2005); in situations like this, the probability of nonresponse depends on measurements which were not collected. The reason(s) for missing data must be explicitly modeled, often based upon untestable assumptions. Thus, there is an acute need for additional methods for handling missing data that require fewer assumptions about missing data.

The frequency of missing data can be surprisingly high; Kim et al. (2011) found newly added variables to the SEER registry where at least 50% of cases had missing data. In a meta-analysis of intensive care clinical studies, Vesin et al. (2013) noted less than 5% of the studies used a sophisticated technique to handle missing data and 45% ignored records with missing data and conducted a complete case analysis. In the studies which used complete cases, the estimates and standard errors differed by as much as 60% when this team reran the analysis using a more sophisticated technique. Both recent studies, the works of Kim and Vesin are indications that missing data are frequently a challenge when conducting clinical research.

Treating missing data in clinical research has additional elements which require consideration. For instance, most interventions are assumed to improve the outcome of interest (e.g. pain relief in a study of analgesics) but sometimes the goal of intervention is to delay disease progression. In a lengthy trial which aims to limit the progression of Alzheimer's disease, applying LOCF imputation to dropouts assumes no further worsening of their condition. This assumption is generally incorrect and could lead to

making false conclusions regarding the effectiveness of the intervention or the natural progression of the disease.

Motivation for the Present Analysis

The purpose of this work is to identify if a candidate technique performs as well or better than current approaches in the presence of MNAR data. Non-ignorable missing data are recognized as an issue that reduces the quality of statistical inferences when they are not handled appropriately. Due to the intrinsic properties of MNAR data, the techniques which can handle this type of missing data carry stricter assumptions and are more difficult to model compared to techniques for ignorable data.

The primary aim is to optimize the tuning parameters of the candidate technique to the chosen data set. Every data set requires a small but appreciable amount of fine tuning before applying this technique. The secondary aim is to compare the performance of the candidate technique to conventional approaches. These approaches are ones likely to have been used as alternatives to the candidate and are varied in their expected performance for treating MNAR data. These aims, broken down into more specific research questions in Chapter 3: Methods, should provide an appraisal of how well the candidate technique is able to handle non-ignorable missing data.

CHAPTER 3: METHODS

Dynamic Cluster-Based Imputation

Dynamic cluster-based imputation (DCI) was selected as the potential candidate for addressing non-ignorable data. Cluster-based sampling approaches follow a simple philosophy: the best estimate of a missing value comes from evaluating the most similar cases with non-missing values. Early clustering techniques focused on splitting a data set into a discrete number of clusters (Mantras, 1991) and imputing missing values based on the local information from the members of the cluster. With no influence from noisy data outside the cluster, demographically similar cases should result in more accurate predictions. The primary concern with this methods was that creating discrete clusters resulted in subgroups that were too fine (similar cases were isolated across many clusters) or too coarse (too many dissimilar cases were constrained by too few clusters). Ayuyev et al., (2009) proposed dynamic cluster-based imputation which relies on the similarity of information from an element's shared neighbors. Dynamic clustering uses a draw-and-replace method of selecting most similar neighbors; any element can be declared a neighbor to any number of observations. In effect, each observation has its own cluster of similar cases instead of parsing drawn-without-replacement clusters.

DCI has the potential to accurately predict missing values from non-ignorable missing data. Using the local information of similar cases via cluster membership is the main strength of this method, along with requiring only two user-specified parameters which reduces the tuning necessary before running. Leveraging local information from the data set to predict missing values eliminates the need to model an explicit missing

data mechanism; the kind that would be necessary during a sensitivity analysis. Since the profile of each data set is unique, identifying the clustering pattern that best fits the shape, size and density is challenging (Khani et al., 2013).

Generation of Simulated Data

Instead of using actual research data, the data set used in this work has been simulated. The benefit of using simulation data is that the true mechanism responsible for missing data is known because it is chosen as part of the simulation's design. "Real" data are never so neat and precise, making it impossible to know for certain the pattern of missing data. The simulated design comes from a report on addressing missing data in cluster-randomized trials from Puma et al. (2009); based on a hypothetical cluster-randomized education trial intended to measure the impact of an intervention on student achievement, randomization takes place at the school level. With equal likelihood schools are assigned to either an unspecified treatment or the control which was the lack of an intervention.

Specifically, these features of the Puma et al. design are carried over into the simulated data of this work:

- One sample contains 60 schools with 30 schools assigned to either arm.
- Each school includes 60 students (cases), all of whom receive their treatment determined at the school level.
- The following baseline factors are known for all cases: gender, a high risk status variable and their pre-test assessment scores.

With only these variables, one thousand replicate samples were simulated to generate pre-test and post-test scores for all cases.

Pre-test scores for all cases were generated from the following model:

$$Y_{Pre,ij} = \beta_0 + \beta_1(\text{FemaleCentered}) + \beta_2(\text{HighRiskCentered}) + \alpha_{0j} + \varepsilon_{ij}$$

$\beta_0 = 0$
$\beta_1 = 0.20$
$\beta_2 = -0.8$
$\alpha_{0j} \sim N(0, 0.1)$
$\varepsilon_{ij} \sim N(0, 0.9)$

Post-test scores for all cases were generated from the following model:

$$Y_{Post,ij} = \beta^*_0 + \beta^*_1(\text{FemaleCentered}) + \beta^*_2(\text{HighRiskCentered}) + \beta^*_3(Y_{Pre,ij}) + \beta^*_4(Trt_j) + \beta^*_5(Trt_j \times Y_{Pre,ij}) + C(\alpha^*_{0j} + \varepsilon^*_{ij})$$

$\beta_0 = 0$
$\beta^*_1 = 0.02$
$\beta^*_2 = -0.05$
$\beta^*_3 = \sqrt{(r_{Pre,Post})}$
$\beta^*_4 = 0.20$
$\beta^*_5 = -0.20/3$
$C = \sqrt{(1- r_{Pre,Post})}$
$\alpha^*_{0j} \sim N(0, 0.1)$
$\varepsilon^*_{ij} \sim N(0, 0.9)$

In most education settings, students tend to be more similar to other students in their school than to students in other schools. As a result, some of the variability in student achievement can be explained by variability across schools. This study assumed an intraclass correlation (ICC) of 0.10 in pretest scores, which means that 10 percent of the variation in achievement across students can be explained by knowing mean pre-test scores of the school. Puma et al. added some realistic qualities to the scores as well based on research collected from public school assessments conducted in California. Average

pretest scores are 0.20 standard deviations higher for girls than for boys; additionally, average pretest scores are 0.80 standard deviations lower for high-risk students than for low-risk students. Treatment had an average effect size of 0.20, meaning that students in the intervention arm would have an average benefit to their post-test scores of 20 percent of a standard deviation compared to control group students.

After pre- and post-test scores were generated for every case, missing post scores were determined by masking a specific percentage of those values based on an interaction between treatment group and post-test scores in order to simulate MNAR missing data. Missing data were more likely for students with lower post-test scores than for students with higher post-test scores. In addition, the average rate of missing data was higher in the schools assigned to the control group than the schools receiving the intervention. As noted by Wolf et al. (2009), this difference is frequently due to greater cooperation with data collection by treatment group schools and students compared to control schools and students. Overall, the rate of missing post-test scores was 40%, but it could be considerably higher or lower in specific strata of the data. Broken down by treatment, cases in the intervention arm were missing on average 35% while control arm cases were missing 45%. The probability of nonresponse was a function of treatment status and post-test score as noted in Table 1.

Table 1: Missing Data Probabilities Used to Generate the Data Set.

Post-test Quartile	Treatment	Control
1 (Highest)	30	30
2	35	40
3	35	50
4 (Lowest)	40	60
Average	35	45

Research Questions

The purpose of this study is to investigate the performance of dynamic cluster-based imputation relative to other missing data techniques. MNAR missing data is the most challenging type of missing data to accurately recover parameters from. Many popular strategies reach biased conclusions when data are MNAR (e.g., multiple imputation and full information likelihood) and the unbiased methods require explicit modeling of untestable missing data mechanisms (e.g., pattern-mixture and Diggle-Kenward selection models). Therefore, there is a need for an imputation strategy for non-ignorable data that does not require explicit modeling of a missing data mechanism. Using local information instead of modeling the missingness mechanism, DCI is a technique which, if successful, could meet this criterion.

In this first experiment, the performance of DCI was evaluated in comparison to two simplistic, but widely employed strategies that were not designed for use with MNAR data (listwise deletion [LW] and last observation carried forward [LOCF]) along with the two most commonly used strategies for handling non-ignorable data (pattern-mixture models [PM] and Diggle-Kenward [DK] selection models). This first study was designed to address the following three research questions:

1) What combination of nearest neighbors and cluster size is optimal for accurately imputing MNAR missing data in this cluster-randomized design?

2) Based on the optimal parameters identified by the first research question, how well does DCI perform when compared with other conventional missing data techniques (LW, LOCF, PM, and DK)?

3) Does increasing the number of nearest neighbors and cluster size improve results compared to the smaller number of nearest neighbors from the exploratory analysis?

Outcome Measures

Four primary outcomes were used to measure the success of the chosen methods: estimated treatment effect (b), standard error of estimated treatment effect (SE(b)), root mean squared error (RMSEA) and coverage rates (COVG). Treatment effect was a fixed coefficient in the data simulation models. Representing the change between pre- and post-test scores attributable to the treatment, the true value of the treatment effect was 0.20. Standard error estimates the uncertainty of the sampling distribution. The root mean squared error measures the difference between predicted values and observed values and serves as a gauge for relative model quality comparisons. Coverage rates measure the accuracy of predicted treatment effects relative to the true treatment effect. When the 95% confidence interval of a predicted treatment effect includes 0.20 it is considered a success; the percentage of successes is noted as the coverage rate.

DCI relies on a distance measurement that is able to consider both categorical and continuous variables. Using the notation dist(x_i, x_j), the distance between any two cases is calculated with Gower's (1971) coefficient, which ignores missing values in the

calculation of distance measures. After standardizing each variable, the similarity of the cases is calculated as the summed distance of all the individual variable distances. In this data, the variables gender, high risk status and pre-test achievement determined the similarity of cases. The results of every pairwise calculation of Gower coefficient are condensed into an [N x N] matrix, where N is the number of cases in the data set.

[Dist] =

Table 2: Example Calculation of Gower Coefficients.

case	1		…		N
1		$dist(x_1, x_2)$	…	…	$dist(x_1, x_N)$
	$dist(x_2, x_1)$				$dist(x_2, x_N)$
…	…				…
	…				…
N	$dist(x_N, x_1)$	$dist(x_N, x_2)$	…	…	

Mathematically, the most similar cases are those with the smallest Gower coefficient, indicating the smallest summed distance across all variables. With an ascending sort, each row provides a sequential listing of every distance from smallest to largest.

The first user-specified tuning parameter, number of nearest neighbors (K) is now utilized to define a neighborhood of potential cases to be considered in imputation. After all rows have been sorted from smallest to largest distances, the nearest neighbors are the K cases with the smallest distances. Effectively, this reduces the size of [dist] to the K left-most columns instead of N columns. Of interest now are the j^{th} case of each (i, j) pair.

For example, the matrix of K nearest neighbors [KNN] could have cells filled in a manner similar to this:

[KNN] =

Table 3: Example Filled K Nearest Neighbors Matrix.

case	Smallest distances, sorted from smallest to largest				
1	25	62	10	...	12
2	3	97	68	...	51
3	97	2	68	...	105
…					
N	18	12	25	...	33

In this example, case 1 is most similar to cases 25, 62, 10,... and least similar to case 12. The numeric distances are not necessary but will be used again in a future step.

With the K nearest neighbors defined for each case, the next step is to create the neighborhood matrix [NM]. The neighborhood matrix counts the number of nearest neighbors shared in common between every pairwise combination of cases. For example, in cell [2, 3] (or [3, 2] because only calculations either above or below the main diagonal are necessary) a 2 indicates the two shared neighbors: observations 68 and 97. All cells are filled with the same [i, j] computation but are left blank here to help illustrate the single worked example.

[NM] =

Table 4: Example of a Single Cell Calculation in the Neighborhood Matrix.

case	1	2	3	…	K
1					
2			2		
3		2			
…					
N					

Larger values in [NM] indicate more shared neighbors between those two cases. Two cases which share many neighbors are likely more similar than two cases which share no neighbors.

The penultimate matrix necessary for dynamic cluster-based imputation [DCI] calls upon the information from [Dist] and [NM]. Each cell of [DCI] is calculated as the ratio of Gower coefficient to number of shared neighbors:

$dist(x_i, x_j) / [NM_{i,j}]$

[DCI] =

Table 5: Example of a Single Cell Calculation in the Final Matrix.

case	1	2	3	…	K
1					
2			$dist(x_2, x_3)/2$		
3					
…					
N					

When cases are similar, their dist(x_i, x_j) will be small. When cases share many neighbors their [$NM_{i,j}$] count will be high, relative to K. The cells of interest are the ones which have both small distances and high counts of shared neighbors, resulting in the smallest [DCI] values. With an ascending sort, each row provides a sequential listing of every ratio from smallest to largest. The second user-specified tuning parameter, cluster size (R) is now utilized. After all rows have been sorted from smallest to largest distances, the cluster size are the R number cases with the smallest ratio. Effectively, this reduces the size of [DCI] to the R left-most columns instead of K columns. Similar to finding the K nearest neighbors based on the Gower coefficients, of interest again are the j^{th} case of each (i, j) pair. For example, the matrix of cluster size R [List] could have cells filled in a manner similar to this:

[List] =

Table 6: Example of a Filled Cluster Membership Matrix.

case	Smallest ratios, sorted from smallest to R^{th} smallest				
1	10	62	25	...	18
2	3	47	61	...	51
3	2	20	68	...	6
…					
N	12	31	18	...	99

Each row of [List] contains the cluster membership for that case. Cluster membership for case 1 are 10, 62, 25,... 18. These are the R smallest ratios of Gower coefficient to number of shared neighbors from [DCI]. All members of this cluster who have a non-missing value on the variable being imputed are used for the case's

imputation; their relative contribution toward the imputed value is weighted by their value in the neighborhood matrix [NM]:

$$x_{i,j}^{ms} = \left[\sum_{r=1}^{R} x_{r,j}^{(C_{i,j})} nm_{i,r} \right] \Big/ \sum_{r=1}^{R} nm_{i,r}$$

CHAPTER 4: RESULTS

Experiment 1

Experiment 1 was performed to discover the optimal combination of nearest neighbors(K) and cluster size (R) from a wide selection of parameter combinations. The DCI algorithm is highly computationally-intensive, with order of operations $O(N^3 \log N)$, which meant that running the algorithm across a large number of conditions for the full sample and a large number of replications would take prohibitively long to run. For this reason, all combinations in Experiment 1 were run using 1000 replications on a sub-sample of 8 schools (1.33% of the total data).

Design

Combinations of {10, 20, 30, 40, 50, 60} nearest neighbors and cluster sizes {1/5, 1/4, 1/3, 1/2, 1} of nearest neighbors were considered in a 6 × 5 factorial design.

Cluster sizes (R) were determined with the simple algorithm:

$$R = \text{integer}(K/i) \qquad \text{where } i = \{5, 4, 3, 2, 1\}$$

The combinations of parameters used in Experiment 1 are noted in Table 7.

Table 7: Nearest Neighbor and Cluster Sizes Investigated During Experiment 1.

Number of Neighbors (K)	Cluster Size (R)
10	2*, 3, 5, 10
20	4, 5, 6, 10, 20
30	6, 7, 10, 15, 30
40	8, 10, 13, 20, 40
50	10, 12, 16, 25, 50
60	12, 15, 20, 30, 60

* The second instance of 2 (integer(10/5) = integer(10/4) = 2) is omitted.

Estimated Treatment Effect

Estimated treatment effect was under-estimated, by approximately 50%, in every combination of nearest neighbors and cluster size and was largely unaffected by number of nearest neighbors. Interestingly, the least biased treatment effect was obtained for a cluster size of 10 observations for all neighborhood sizes except 60 (the number of observations per cluster/school), where it was 12, the smallest cluster size considered (1/5 of neighborhood size).

Root Mean Square Error

Overall, root mean square error (RMSE) increased as the number of neighbors increased. Similarly, in general, the RMSE increased as the cluster size increased, except in the smallest cluster size where it decreased until all 10 neighbors were considered. RMSE was lowest for cluster sizes of 10, 10, 7, 8, 10, and 12 for numbers of neighbors 10, 20, 30, 40, 50, and 60, respectively. Thus, RMSE favored imputation based on a small cluster of neighbors of approximately 10 observations.

Standard Error

Standard errors of estimates followed a pattern similar to RMSE. Specifically, standard errors were generally larger for larger numbers of neighbors, and as cluster sizes increased. For all numbers of neighbors except the smallest (10), the standard error was smallest for cluster sizes of 10. With 10 neighbors, the standard error was smallest with cluster sizes of 2 and 3 (0.094), but only slightly larger with cluster sizes of 5 and 10 (0.095).

Coverage Rate

Coverage rates were very stable across combinations of number of neighbors and cluster sizes. However, they tended to favor smaller cluster sizes for all number of neighbors except 40, where the highest coverage rates were obtained when all neighbors were used for imputation.

Results across each metric used to evaluate the quality of imputations tended to favor smaller cluster sizes, with 10 observations being at or close to the optimal values in each condition. Results were much more sensitive to cluster size than number of neighbors. Thus, the combination recommended by Ayuyev and colleagues (2011) were selected for experiment 2 (K=50; R=9). Results for Experiment 1 are shown in Table 8 and Figures 2, 3, 4, and 5 below.

Table 8: Results of Experiment 1 by Nearest Neighbors (K) and Cluster Size (R).

Nearest Neighbors (K)	I index	Cluster Size (R)	Treatment Effect	RMSE	Standard Error	Coverage Rate
10	5,4	2	0.096	0.316	**0.094**	**73.0%**
	3	3	0.095	0.301	**0.094**	72.0%
	2	5	0.094	0.302	0.095	71.0%
	1	10	**0.097**	**0.288**	0.095	72.0%
20	5	4	0.099	0.295	**0.096**	72.0%
	4	5	0.099	0.293	**0.096**	73.0%
	3	6	**0.100**	0.287	**0.096**	**74.0%**
	2	10	**0.100**	**0.286**	**0.096**	**74.0%**
	1	20	0.091	0.336	0.098	72.0%
30	5	6	**0.098**	0.292	**0.097**	**73.0%**
	4	7	**0.098**	**0.290**	**0.097**	**73.0%**
	3	10	**0.098**	0.292	**0.097**	**73.0%**
	2	15	0.094	0.315	0.098	**73.0%**
	1	30	0.075	0.385	0.101	66.0%
40	5	8	**0.098**	**0.288**	**0.097**	71.0%
	4	10	**0.098**	0.290	**0.097**	73.0%
	3	13	0.096	0.300	**0.097**	72.0%
	2	20	0.088	0.342	0.099	73.0%
	1	40	0.096	0.364	0.101	**76.0%**
50	5	10	**0.100**	**0.290**	**0.096**	74.0%
	4	12	0.099	0.296	0.097	**75.0%**
	3	16	0.095	0.315	0.098	73.0%
	2	25	0.086	0.367	0.101	73.0%
	1	50	0.087	0.369	0.101	71.0%
60	5	12	**0.102**	**0.295**	**0.097**	**75.0%**
	4	15	0.100	0.309	0.098	74.0%
	3	20	0.094	0.341	0.099	73.0%
	2	30	0.092	0.357	0.101	74.0%
	1	60	0.072	0.376	0.099	65.0%

The optimal value(s) of each measure in each condition are shown in bold.

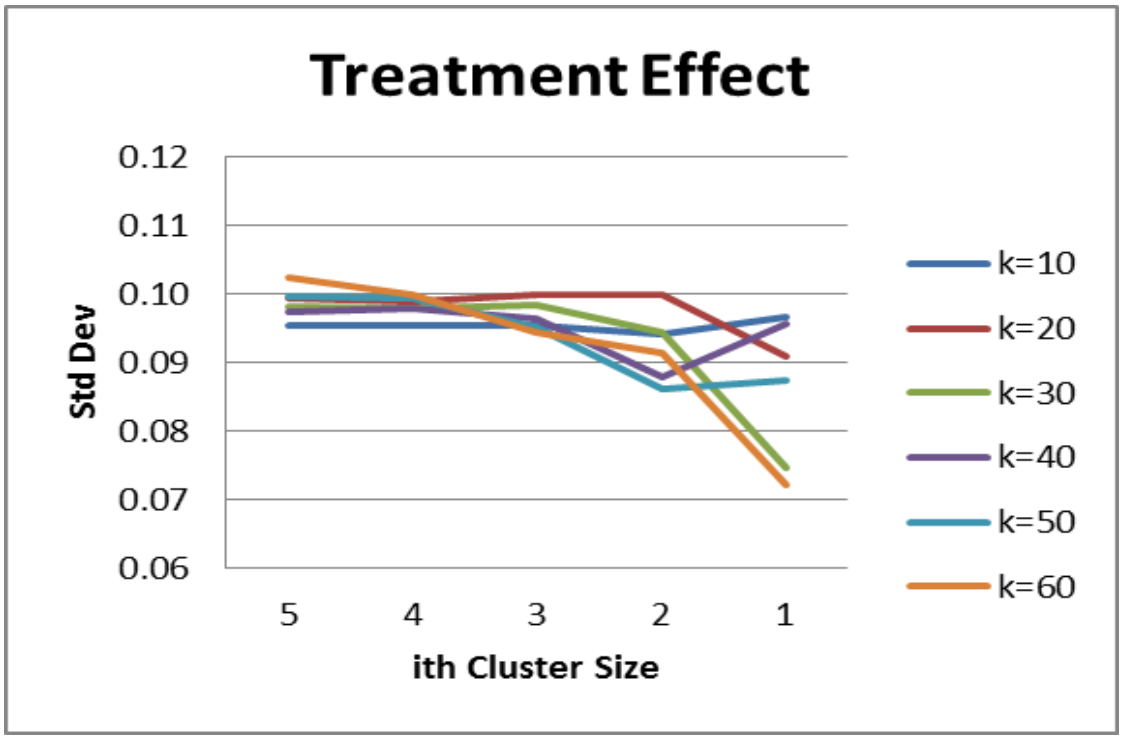

Figure 2: Treatment Effects Observed During Experiment 1.

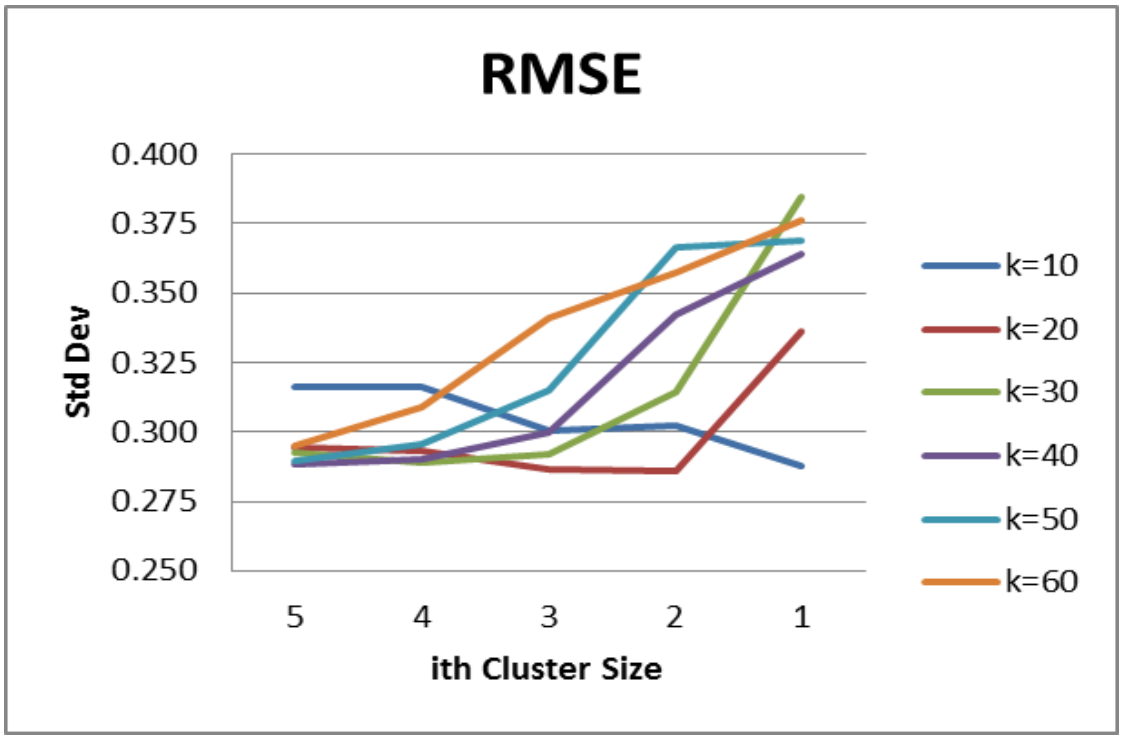

Figure 3: RMSEs Observed During Experiment 1.

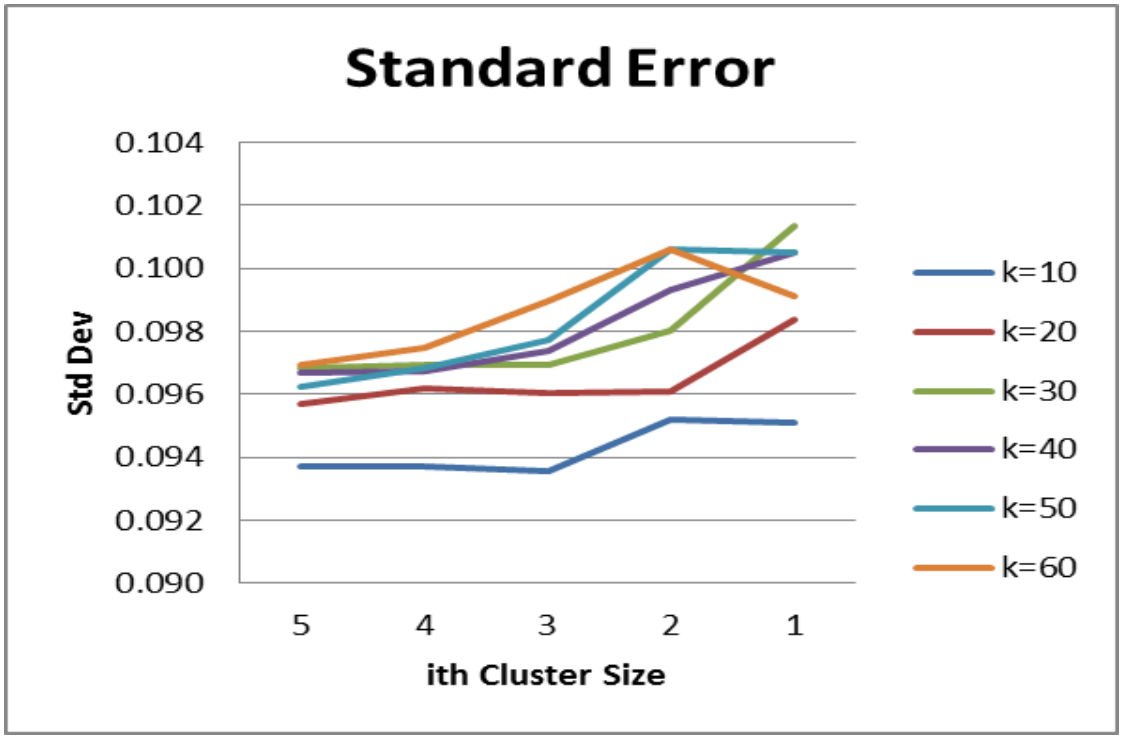

Figure 4: Standard Errors observed During Experiment 1.

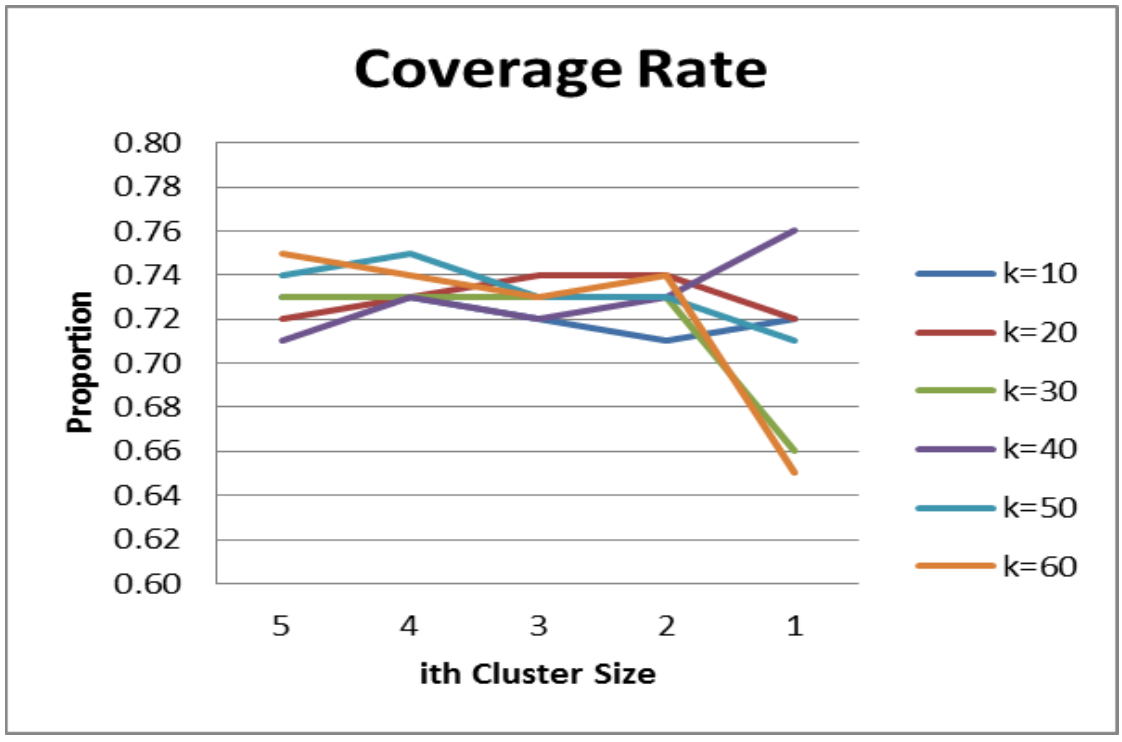

Figure 5: Coverage Rates Observed During Experiment 1.

Experiment 2

Having determined an appropriate combination of number of neighbors (K) and cluster size (R) from Experiment 1, the purpose of Experiment 2 was to evaluate the performance of DCI with listwise deletion (LW), last observation carried forward (LOCF), pattern-mixture models (PM), and Diggle-Kenward (DK) selection models to analysis of the full data set. 1000 replications of 60 schools (100% of the data) were considered. Results for all five techniques are presented in Table 9 below.

Table 9: Results of DCI (K=50; R=9) and Comparison Techniques.

Technique	Treatment Effect	RMSE	Standard Error	Coverage Rate
DCI (K=50; R=9)	0.091	0.290	0.046	33.2%
Listwise Deletion	0.129	0.075	0.051	70.8%
LOCF	0.111	0.089	0.032	18.5%
Pattern-Mixture	0.132	0.072	0.050	71.4%
Diggle-Kenward	0.128	0.073	0.050	70.8%

Surprisingly, DCI performed substantially more poorly for the full data set than in Experiment 1 on all measures. DCI also performed substantially more poorly for all outcomes than did any of the other techniques.

Estimated Treatment Effect

In terms of estimated treatment effect, the best results were obtained from pattern-mixture models (0.132), followed by listwise deletion (0.129), Diggle-Kenward (0.128), last observation carried forward (0.111), and DCI (0.091).

Root Mean Square Error

For RMSE, results followed a similar pattern. The smallest RMSE values were found for pattern-mixture models (0.072), followed by Diggle-Kenward models (0.073),

listwise deletion (0.075), last observation carried forward (0.089), and distantly DCI (0.290).

Standard Error

Standard errors were fairly similar among the five methods. LOCF had the lowest (0.032) followed by DCI (0.046) while pattern-mixture models, Diggle-Kenward models and listwise deletion had the highest standard errors (0.050, 0.050 and 0.051, respectively).

Coverage Rate

In terms of coverage rates, pattern-mixture models had the best coverage rates (71.4%) followed by Diggle-Kenward models (70.8%) and listwise deletion (70.8%), DCI (33.2%), and last observation carried forward (18.5%).

Overall, DCI performed very poorly in analyses of the full data set. These findings were quite unexpected and differed substantially from the results obtained from Experiment 1. For all measures except standard error (where all techniques performed similarly), DCI performed most poorly or second most poorly, with results worse than listwise deletion. Because these results were unexpected and divergent from expectations, a third experiment was conducted in order to consider a wider range of combinations of number of neighbors and cluster size in case the larger sample size in Experiment 2 affected the optimal combination or tuning parameters.

Experiment 3

The purpose of Experiment 3 was to consider a wider range of combinations of number of neighbors and cluster size in order to determine whether a different combination of tuning parameters was required for analyses with the full sample. Factorial combinations of number of neighbors K = {60, 120, 180, 24} × cluster size R = {10, 30} were considered with 100 replications of 60 schools (10% of the data) in each condition. The combinations of tuning parameters for Experiment 3 are shown in Table 10 below.

Table 10: Nearest Neighbor and Cluster Sizes Investigated During Experiment 3.

Number of Neighbors (K)	Cluster Size (R)
60	10, 30
120	10, 30
180	10, 30
240	10, 30

Results for Experiment 3 are shown in Table 11 and Figures 6, 7, 8, and 9 below. They show a very clear pattern. Specifically, considering larger numbers of neighbors and cluster sizes did not improve upon the performance of DCI under these conditions. Thus, it seems unlikely that the performance could be substantially improved through selection of different tuning parameters.

Table 11: Results of Experiment 3 by Nearest Neighbors (K) and Cluster Size (R).

Nearest Neighbors (K)	Cluster Size (R)	Treatment Effect	RMSE	Standard Error	Coverage Rate
60	10	0.092	0.289	0.047	32.0%
	30	0.070	0.373	0.049	27.0%
120	10	0.094	0.285	0.046	37.0%
	30	0.070	0.335	0.046	23.0%
180	10	0.094	0.286	0.045	37.0%
	30	0.057	0.333	0.041	10.0%
240	10	0.090	0.281	0.045	36.0%
	30	0.056	0.312	0.040	10.0%

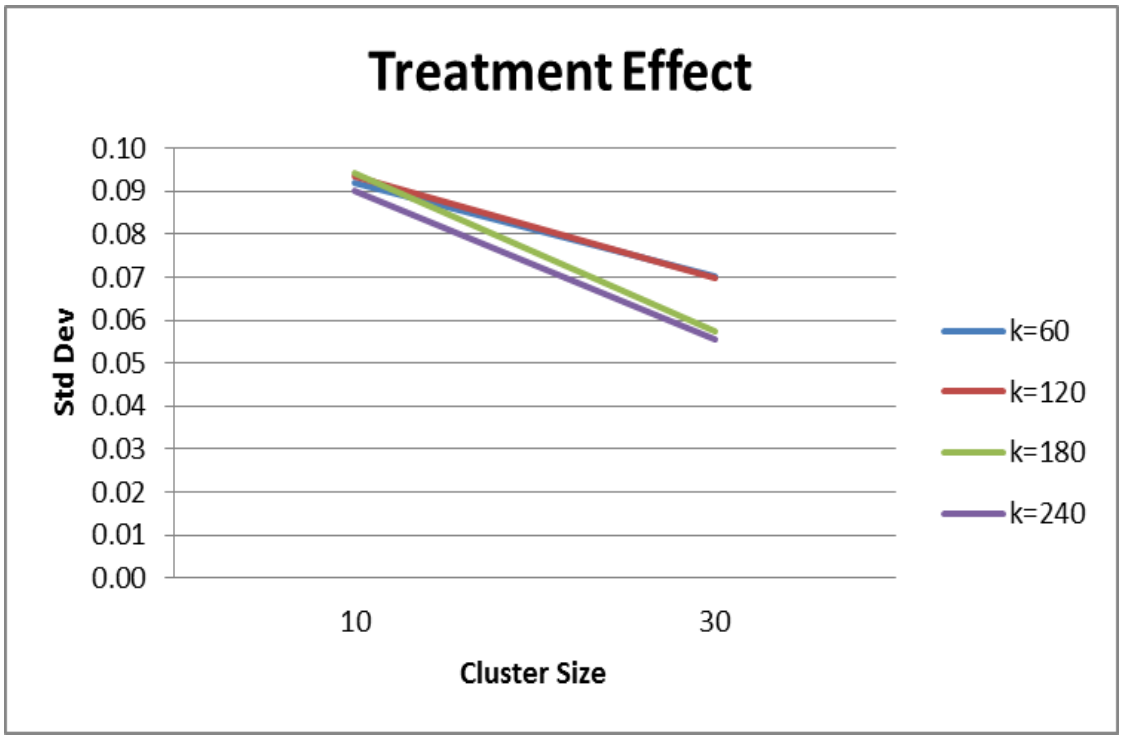

Figure 6: Treatment Effects Observed During Experiment 3.

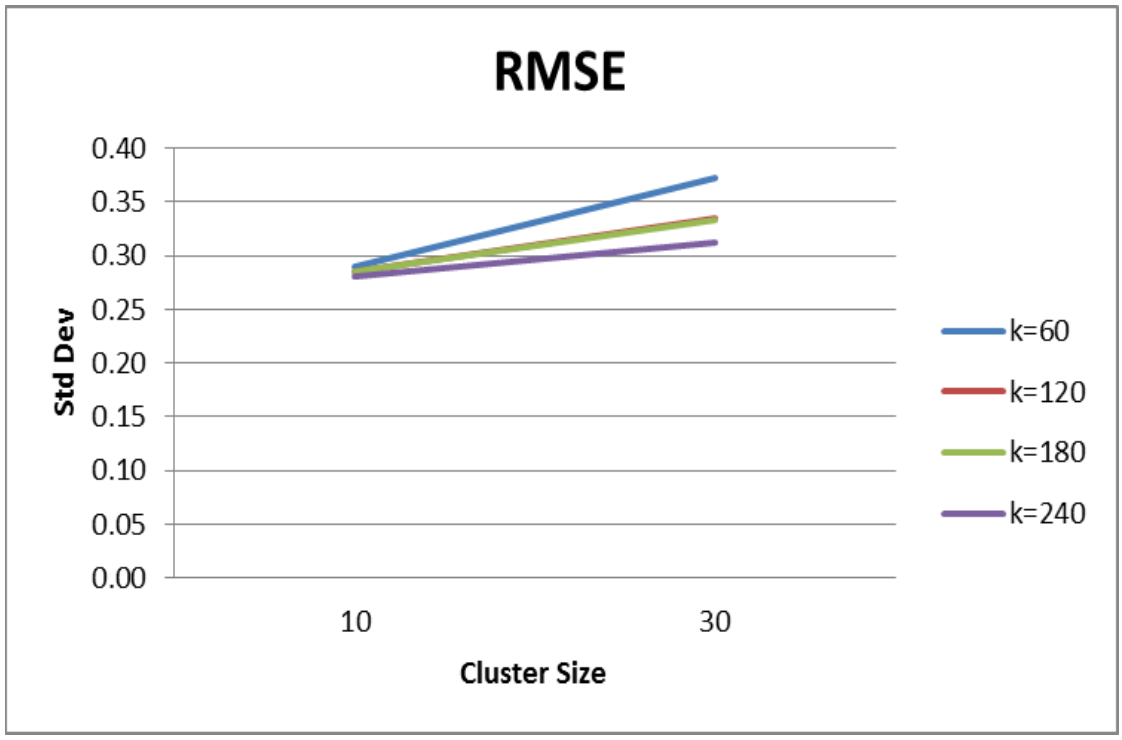

Figure 7: RMSEs Observed During Experiment 3.

Figure 8: Standard Errors Observed During Experiment 3.

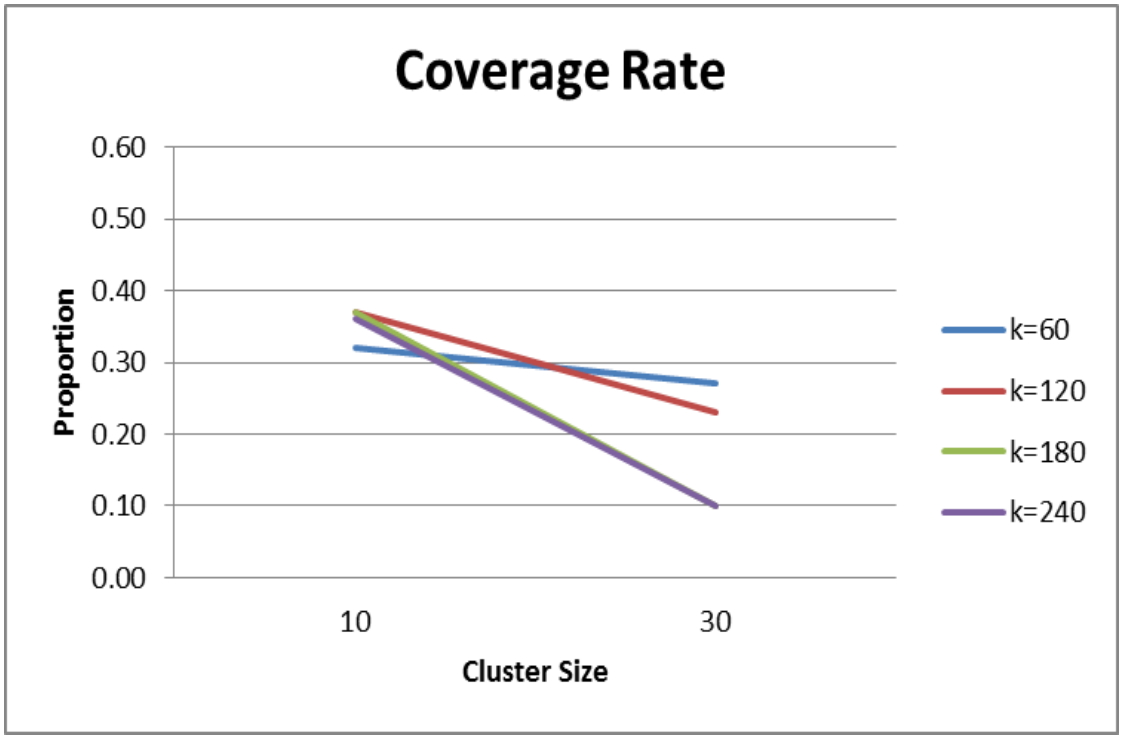

Figure 9: Coverage Rates Observed During Experiment 3.

Verification of the DCI Code

In order to ensure the results observed during the experiment truly represent the capabilities of DCI under MNAR conditions and are not due to coding errors, the same routine was run under additional conditions and those results are briefly discussed in this section.

DCI on MAR Data

By manipulating which variables are used to determine missingness of post-test scores, any of the three types of missing data (MCAR, MAR or MNAR) can be simulated. In the case of the original experiments, MNAR missing data were generated by using the treatment received and post-test score to determine the probability of nonresponse. MAR missing data were also simulated by using only treatment received to determine nonresponse probability. This pattern of missingness for post-test scores can be wholly explained by the treatment received, which is observed for all students, thus making the pattern of missing data MAR. Simulated at low (5%) and high (40%) levels of missing data, these two scenarios will make excellent candidates for ascertaining if the DCI code is performing as expected. The decision to use MAR data is easy to justify; as ignorable missing data, MAR missingness is less challenging to impute which means the quality of the results should improve. Table 12 contains the results of DCI applied to two MAR data sets.

Table 12: Results of DCI on MAR Data at Low (5%) and High (40%) Levels of Missing Data.

Type	Percent Missing	Treatment Effect	RMSE	Standard Error	Coverage Rate
MAR	5%	0.1999	0.311	0.047	93.9%
MAR	40%	0.184	0.277	0.046	91.8%

At both levels of missing data, the DCI routine demonstrates very high imputation accuracy. Treatment effect, fixed at 0.2 in this simulation, is predicted better with DCI on MAR data than any technique from Experiment 2; 0.1999 in the case of 5% missing and 0.184 for 40% missing. Regarding the error terms, RMSE is observed to be higher than the results observed during Experiment 2 but lower than Experiment 3 (0.311 and 0.277 for low and high missing rates, respectively) while these standard errors are comparable to the results of Experiment 2. Coverage rates are also improved over all Experiment 2 techniques, 93.9% in the case of 5% missing and 91.8% for 40% missing. The dramatic improvement in results acts as proof the DCI routine is performing as expected.

Manual Review of DCI Procedure

Another effort to ensure DCI was performing as expected was to manually recreate the matrix [NM] of a much smaller sample (n=30, K=10) and compare it to the coded output of the same sample. Defined in Chapter 3, [NM] calculates the number of shared neighbors between any two cases. Starting from [KNN] it is possible to manually tally the number of shared neighbors for all *i*,*j* cases. Due to the nature of manually computing [NM] no output is provided but there were no discrepancies between manual and coded outputs on this sample. Together, these two validations of DCI serve as proof the coded version of DCI performed as expected.

DCI on MNAR With Additional Missing Data Indicator

A final exploratory analysis with DCI was conducted on the MCAR data set. This analysis introduced an additional covariate: a dummy variable used to indicate whether the post-test score for a student was missing or observed. The benefit of including this missing score indicator is explained by the nature of the pattern of missingness in MNAR data. Recall Table 1 which illustrates the relationship between the value of post-test scores and their rate of being masked as missing. For both treatment and control, lower raw scores were more likely to be missing; this is the fundamental mechanism responsible for the MNAR pattern of missingness present in the data. Of interest is whether the missing data indicator acts strongly enough as a proxy for lower scores which, when acting in concert with the other variables in the model, improves the overall imputation quality.

Table 13: Results of DCI on Original MNAR Data With Addition of Post-Test Missing Dummy Variable as Covariate.

Type	Percent Missing	Treatment Effect	RMSE	Standard Error	Coverage Rate
MNAR	40%	0.128	0.288	0.070	82.8%

As noted in Table 13, inclusion of the missing value indicator improves the results of this analysis when compared against the techniques of Experiment 2, especially the DCI model. The treatment effect of 0.128 rivals the best results observed in Experiment 2 and improved on the previous DCI result (0.091) by forty percent. RMSE (0.288) remained similar between DCI results and was much higher than every other technique. Standard error, for which DCI ranked among the highest in Experiment 2 (0.046), increased by over fifty percent to 0.070. Coverage rate increased greatly with the

additional covariate, climbing from 33.2% to 82.8%. Overall, it is clear that the inclusion of a missing variable indicator improves the accuracy of DCI with notable gains in treatment effect and coverage rate while suffering an increase in standard error.

CHAPTER 5: DISCUSSION

The purpose of this thesis was to evaluate the potential of a new imputation method for use with cluster-randomized data when data are non-ignorably missing, a situation that is difficult to remediate effectively using existing missing data methods. The central hypothesis guiding this work was that, under situations where probability of non-response depends on unobserved variables, imputations based on a local neighborhood of cases similar to the ones with missing data might out-perform methods based upon all cases.

This is an important problem for several reasons. First, the most widely used methods for analysis with missing data, full information maximum likelihood and multiple imputations, do not apply to the realistic and general situation where the probability of missing data depends on unobserved values (i.e., MNAR). Second, existing methods for analysis of MNAR data such as pattern-mixture models and Diggle-Kenward models can be highly sensitive to the explicit modeling assumptions made about missing data mechanisms. Thus, there is a critical need for missing data methods that are effective and easy to tune, but more robust in the face of MNAR data.

While each experiment was designed to answer a specific research question, there were observable trends in the results that appeared across all experimental conditions. Every combination of nearest neighbors and cluster size underestimated the known treatment effect of 0.20. Of all combinations, K=60 with R=12 gave the closest estimate of treatment effect of 0.102. Comparatively, both techniques which model data missingness, the pattern-mixture model and the Diggle-Kenward selection model

estimated the treatment effect as 0.132 and 0.128, respectively. While these estimates may all seem far from the true value, it should be noted that underestimating treatment effect is more desirable than over-stating the benefit provided by the intervention. In their original simulation study, Puma et al. (2009) defined estimates that differed by more than .05 from the original value to represent unacceptable levels of bias, and pointed to the condition examined in the current set of studies as one for which acceptable methods for handling missing data were not available. By this criterion, none of the methods considered here provided results with "acceptable" levels of bias.

Experiments 1 and 3 focused on comparing combinations of tuning parameters. In both instances, strong trends relating to what qualities provide the most favorable results occurred. Optimal conditions, indicated by best estimated treatment effect and high coverage rates with low error terms, routinely occurred when cluster sizes were smallest. Between the two tuning parameters, cluster size appeared to make more of a difference in the resulting imputations than nearest neighbors. Knowing this can reduce the number of potential tuning parameter combinations and thus the time spent running future preliminary experiments.

Experiment 1

Experiment 1 was conducted to determine the optimal combination of tuning parameters for these cluster-randomized data. There was no literature available on the use of dynamic cluster-based imputation on a cluster-randomized design, so it was necessary to conduct an exploratory analysis to gain insight into the optimal parameterization for

this novel application. Justified below, in this instance the optimal combination of parameters was defined as K=50 and R=9.

When studied with MAR missing data, Ayuyev et al. noted that the pairing of K=50 and R=9 was optimal. Similar results were obtained within the range of $40<K<60$ with R=7, 9, or 11. These results paralleled the observations of Experiment 1 because of the parameter combinations studied, K=50 and K=60 tended to best predict treatment effect at i=5 and 4 (the smallest cluster sizes). Though all coverage rates were high ($\geq$65%) the highest coverage rates were also observed with these combinations. In general, the smallest counts of nearest neighbors (K=10 and K=20) showed no real trend that related to their cluster sizes, while the largest counts of nearest neighbors (K=50 and K=60) showed the most pronounced differences in estimation quality across their cluster sizes. All measurements tended to steadily depreciate as cluster size approached the full complement of nearest neighbors; treatment effect and coverage rates decreased while the error terms increased. 60, the maximum value of K in this experiment, was chosen because it matches the 60 subject cluster-randomization design of the data set. Selecting more nearest neighbors than this is guaranteed to include some neighbors outside the case's randomized cohort.

From these results it was concluded that the optimal tuning parameters tends towards the largest counts of neighbors with small cluster sizes. Based on the optimal results of Experiment 1 and supported by the conclusions from Ayuyev, the conditions selected for the full data analysis of Experiment 2 were K=50 with R=9. High coverage rates that indicate treatment effect is well estimated and the fact that DCI requires no

missing data modeling assumptions were promising qualities for the full analysis planned in Experiment 2.

Experiment 2

The second research question was proposed to assess the performance of dynamic cluster-based imputation with techniques commonly used in randomized trials. Four techniques were selected for comparison and run under the same conditions as DCI: listwise deletion, last observation carried forward (LOCF), a pattern-mixture model and a Diggle-Kenward selection model. Listwise deletion and LOCF are the most commonly used and easiest to implement of the comparators; LOCF is the standard method for handling missing data in randomized trials. Pattern-mixture and Diggle-Kenward are specialized models applied when data are suspected to be non-ignorably missing and were expected to perform well, given the simulated MNAR data of all experiments. Both attempt to model the unobserved missing data mechanism which is fundamentally a different approach than the dynamic cluster-based imputation method.

The modest results of the full analysis were not as favorable as expected based on the exploratory analysis from Experiment 1. Predicted treatment effect and RMSE remained similar while the standard error noticeably decreased. While these were all acceptable results, a coverage rate of 33.2% is too low to express confidence in the imputation. This coverage rate can be interpreted as only one of three proposed imputations estimated a treatment effect whose 95% confidence interval included the true treatment effect of 0.2. Therefore, selecting K=50 and R=9 as DCI parameters did not adequately recover treatment effect under the MNAR mechanism with 40% of post-test

data missing. Both techniques that model the missing data mechanism, pattern-mixture and Diggle-Kenward selection, performed well when challenged with this MNAR missing data. Their coverage rates (> 70%) and predicted treatment effects were high while their RMSEs were low. The treatment effects recovered from pattern-mixture and Diggle-Kenward selection models were the two highest seen across all experiments at 0.132 and 0.128 respectively.

The other two techniques, listwise deletion and LOCF, are known to produce biased estimates when missing data are non-ignorable; the direction and magnitude of this bias is unpredictable (Molenberghs et al., 2004). Listwise deletion provided deceptively good results in this case, with an estimated treatment effect of 0.129 and coverage rate above 70%; similar to the two specialized models. Given the mechanism was designed as MNAR it is interesting to note the apparent success of a method that was intended to fail in this context. LOCF met expectations by faring poorly when applied to this data set. Since there are only two periods where measurements are taken, any missing post-test scores were replaced with the case's pre-test score. In effect, cases with missing data were treated as if they experienced no change between the pre-test and post-test assessments, regardless of treatment. The poor performance of LOCF is noteworthy because this is a standard technique used in randomized trials. Although their use has been discouraged by methodologists, listwise deletion and LOCF continue to have widespread use in academic research and randomized controlled trials (Peugh & Enders, 2004).

When compared to DCI, the pattern-mixture model and Diggle-Kenward selection model performed better; LOCF fared poorly while listwise deletion appears to have done an adequate job. Since the results for this combination of neighbors and cluster size indicated DCI underperformed, Experiment 3 was conducted as a supplemental preliminary analysis with new combinations of tuning parameters.

Experiment 3

After unsuccessfully predicting treatment effect with the optimal conditions from Experiment 1, another exploratory analysis was conducted, this time with larger counts of nearest neighbors. As a trend, the optimal results from Experiment 1 tended to be large neighbor counts with small clusters. Experiment 3 leveraged this trend by increasing the count of nearest neighbors while using small fixed cluster sizes.

As illustrated by Figures 6 to 9, all factorial combinations performed similarly in Experiment 3. Like the trend observed in Experiment 1, treatment effect and coverage rates degraded with increased cluster size while RMSE increased. Unlike Experiment 1, the standard errors remained fixed with increased cluster size. Although the results of cluster size 10 outperformed cluster size 30 for all selected neighbor counts, none of these results warranted a full scale analysis. These tuning parameter combinations reported low coverage rates (< 40%), indicating these imputations are not accurate at predicting treatment effect.

The answer to research question three is clear; increasing the number of nearest neighbors does not improve the treatment effect or coverage rate of the results. While the RMSE and standard error may improve, the quality of the imputation is so poor that any

improvement in these values is worthless. Considering that these data are cluster-randomized it comes as little surprise that permitting DCI to use such a high number of nearest neighbors does not improve the results. The count of nearest neighbors in Experiment 1 was limited to 60 so the imputation would not be forced to draw from outside the cluster-randomization pattern in the data set. This was chosen primarily as a theoretical ceiling since it is likely that some cases would find similar neighbors outside their own school in such a large data set.

Among Experiments 1, 2 and 3, nine distinct counts of nearest neighbors were analyzed. The quality of results is lowest at the extreme ends of neighbor counts (K=10 and 240) and are optimal for counts in the middle (K=50 and 60). In all cases better results were observed with small cluster sizes. This leads to the conclusion that the conditions used in Experiment 2 (K=50, R=9) are likely the optimal parameterization of this data. Dynamic cluster-based randomization does not appear to be an appropriate method of handling missing data in this data set, and potentially any cluster-randomized data set. It has been noted by Taljaard, Donner and Klar (2008) that missing data in a cluster-randomized setting requires unique considerations, including addressing the presence of any intracluster correlation, which could be an unexpected shortcoming of DCI. Ayuyev et al. were able to demonstrate the effectiveness of DCI under conditions different from those in this study; it may be that MNAR data from a cluster-randomized design are not suited to this method. Techniques which modeled the missingness mechanism provided satisfactory results when challenged with this MNAR data.

After validating the DCI code with MAR data and with a manual comparison it is easy to come to the conclusion that dynamic cluster-based imputation, leveraging the use of local information from similar cases, is not appropriate in this study design. Though imputation was unsuccessful in this context, there are still findings that can improve the quality of other applications of dynamic cluster-based imputation. It should be noted that DCI performed very well when applied to MAR data, as noted by both Ayuyev and in this document. All methods for handling non-ignorable (MNAR) non-response require explicit specification of a model for the probability of non-response. In contrast, DCI is based on an empirical approach. It is important to note that bias and coverage rates of DCI rivaled those of current model-based approaches. This came at the cost, however, of increased variance. The standard errors and RMSE for DCI were considerably higher than with model-based approaches with MNAR data, but comparable to other methods when data were MAR. However, this does suggest a place for DCI within the growing set of methods for handling MNAR data. Future studies should consider DCI on a non-cluster-randomized design. Other future improvements to DCI include automating the determination of the optimal K and R parameters, potentially using a bootstrapping or jackknife procedure. Data in this work were strictly continuous though this technique could be adapted to handle mixed data sets with discrete variables as well.

Limitations

This study has several limitations which potentially address the varied performance of DCI across the three experiments. The first limitation is the necessary tuning of nearest neighbor and cluster size parameters. Potentially unique to each data

set, identifying the optimal combination of parameters before implementation of DCI will require the researcher's attention. Although the optimal parameterizations between this study and the work of Ayuyev et al. are very similar, there is little prior research on this technique. It cannot be said with certainty that the optimized ranges ($40<K<60$ and $R=7$, 9, or 11) shared by these two studies are common to all data sets. Additionally, the time and resources necessary to run a full DCI analysis dramatically scale by the size of the data; computational demands increase along the order of $N^3 \log N$ which makes running all combinations of interest at full scale unfeasible for large data sets.

The preliminary experiments appear to be highly sensitive to the subsample of cases which are selected. For example, the combination of tuning parameters $K=60$ with $R=30$ appear in Experiment 1 and Experiment 3 with very different results. In Experiment 1 this combination performed strongly with an estimated treatment effect of 0.092 and a coverage rate of 74%; in Experiment 3 the same combination performed poorly with an estimated treatment effect of 0.070 and a coverage rate of 27%. The difference between these two results can be attributed to the population subsample used. This is important to note because it provides an explanation for the discrepancy between the results of Experiment 1 and Experiment 2. By chance, the subsample used in Experiment 1 may not have been representative of the data set and exaggerated the capacity for DCI to recover this data's treatment effect. When applied to the full data set in Experiment 2, the true profile of the data was not as compatible with dynamic cluster-based imputation.

In this data set there are three baseline characteristics observed for all students: gender, high risk status and their pre-randomization test achievement. These three factors determined the similarity between students used for DCI. Many longitudinal studies collect more baseline variables than this; it is plausible that DCI would be more effective given more baseline characteristics than were present in this data set. The inclusion of a missing data indicator, easily added by statistical packages, was noted above as a means of improving the performance of DCI as well.

CHAPTER 6: CONCLUSION

Missing data is an issue every researcher will address during their career, especially in fields with human participants like public health. With untestable assumptions, non-ignorable missing data present a challenge because many standard techniques for handling missing data are prone to biased estimation (Little and Rubin, 2002). Specialized techniques which model the unknown missing data mechanism have gained favor but a method for addressing missing data through another pathway would be a valuable tool. Dynamic cluster-based imputation was a potential candidate for filling this gap; its use of local information avoids any modeling of a missing data mechanism.

Data simulated as a cluster-randomized trial originally proposed by Puma et al. (2009) served as the data set in this study. In cluster-randomized trials treatments are assigned at the level of a social unit such as a school or neighborhood. Techniques which are commonly used to handle missing data, as well as those designed for non-ignorable data were included as comparisons to DCI. The results indicated that DCI was not able to recover treatment effect accurately on these data, though techniques designed for MNAR data fared better.

While not successful on this data set, no technique is superior for handling all types of missing data; the results of this study may be limited by using a method unsuited to the simulated data. A cluster-based imputation technique may not be a better solution to cluster-randomized data than those which are currently favored. There are other applications for DCI where the technique may prove more successful in handling missing data, like in instances of ignorable missing data. There are future directions and

improvements to the dynamic cluster-based imputation method which could be beneficial as well, like automating the selection of optimal parameters. Used under different conditions than those found in this study, DCI can still be a valuable method for handling missing data.

REFERENCES CITED

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.). *Advanced structural equation modeling. 243–277.*

Ayuyev, V. V., Jupin, J., Harris, P. W., & Obradovic, Z. (2009). Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data. *Data Warehousing and Knowledge Discovery, 11th International Conference, DaWaK 2009 Proceedings. 366-377*

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale: L. Erlbaum Associates.

Davey, A., & Savla, J. (2010). Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach. Routledge: Philadelphia.

Diggle, P., Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics. 43, 49-73.*

Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology.60, 549-76.*

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics, 27(4), 858-871.*

Khan, A., Schwartz, K., Redding, N., Kolts, R. L. and Brown, W. A. (2007). Psychiatric Diagnosis and Clinical Trial Completion Rates: Analysis of the FDA SBA Reports. *Neuropsychopharmacology. 32, 2422-2430.*

Khani, F., Javad Hosseini, M., Ali Abin, A., Beigy, H. (2013). An algorithm for discovering clusters of different densities or shapes in noisy data sets. *2013 Proceedings of the 28th Annual ACM Symposium on Applied Computing. 144-149*

Kim, H.M., Goodman, M., Kim, B.I., Ward, K.C. (2011). Frequency and determinants of missing data in clinical and prognostic variables recently added to SEER. *Journal of Registry Management. 38(3), 120-131.*

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data. New York: John Wiley & Sons.

Little, R. J. A. (2008). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (Eds.). *Longitudinal Data Analysis. 409-431.*

Mallinckrodt, C. H., Sanger, T. M., Dube, S., DeBrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z. and Tollefson, G. D. (2003). Assessing and Interpreting Treatment Effects in Longitudinal Clinical Trials with Missing Data. *Society of Biological Psychiatry. 53, 754-760.*

Mantras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning, 6, 81-92.*

Molenberghs, G. (2009). Incomplete Data in Clinical Studies. *Drug Information Journal, 49, 409-429.*

Molenberghs G., Thijs H., Jansen I., Beunckens C., Kenward M.G., Mallinckrodt C., Carroll R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics, 5(3), 445-464.*

National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials.* Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525– 556.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What to do when data are missing in group randomized controlled trials. US Department of Education, Institute of Education Sciences.

Raghunathan, T. E. (2004). What Do We Do With Missing Data? Some Options For Analysis of Incomplete Data. *Annual Review of Public Health, 25, 99-117.*

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

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

Stuart, E.A., Azur, M., Frangakis, C., Leaf, P. (2009). Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American Journal of Epidemiology. 169(9), 1133-1139.*

Taljaard, M., Donner, A. and Klar N. (2008). Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal. 50(3), 329-345.*

Vesin, A., Benbenishty, J., Timsit, J.F., Azoulay, E., Ruckly, S., Vignoud, L., Rusinovà, K., Benoit, D., Soares, M., Azeivedo-Maia, P., Abroug, F. (2013). Reporting and handling missing values in clinical studies in intensive care units. *Intensive Care Medicine, 39(8), 1396-1404.*

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., & Eissa, N. *Evaluation of the DC Opportunity Scholarship Program: Impacts After Three Years.* (2009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.). *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples. 219–240.*

Yang, X., Li, J. and Shoptaw, S. (2008) Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Statistics in Medicine, 27(15), 2826-2849.*

Yang, X. and Shoptaw, S. (2005). Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug Alcohol Dependence, 77(3), 213-225.*