Title

Handling Missing Outcome Data in Cluster Randomized Trials with Both Individual- and Cluster-Level Dropout

Author
Avila, Analissa Danielle

Publication Date
2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Handling Missing Outcome Data in Cluster Randomized Trials with Both Individual- and Cluster-Level Dropout

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Analissa Danielle Avila

2024

ABSTRACT OF THE DISSERTATION

Handling Missing Outcome Data in Cluster Randomized Trials with Both Individual- and Cluster-Level Dropout

by

Analissa Danielle Avila
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2024
Professor Catherine Crespi-Chun, Chair

Missing outcome data are common in cluster randomized trials (CRTs) which can complicate inference. Further, the missingness can occur due to dropout of individuals, termed “sporadically” missing data, or dropout of clusters, termed “systematically” missing data, and these two types of missingness could have potentially different missing data mechanisms. We aimed to develop a well-performing and practical approach to handle inference in CRTs when outcome data may be both sporadically and systematically missing. To this end, we first examined the performance of four multilevel multiple imputation (MI) methods to handle sporadically and systematically missing CRT outcome data via a simulation study. Our findings showed that one multilevel MI method which uses the maximum likelihood estimates obtained from a linear mixed model to draw missing values outperformed the others under various scenarios. Using the best performing MI method, we developed methods for conducting sensitivity analysis to test the robustness of inferences under different missing at random (MAR) and missing not at random (MNAR) assumptions. The methods allow for different

MNAR assumptions for cluster dropout and individual dropout to reflect that they may arise from different missing data mechanisms. We developed graphical displays to visualize sensitivity analysis results. Our methods are illustrated using a real data application.

The dissertation of Analissa Danielle Avila is approved.

Hilary Jeanne Aralis

Thomas R Belin

Beth Ann Glenn-Mallouk

Catherine Crespi-Chun, Committee Chair

University of California, Los Angeles

2024

To my parents, who have encouraged and supported me always

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Kate Crespi, who has been an incredible mentor. She has spent many hours discussing and reviewing this work with me over the years. From her I have learned to think critically about statistical problems, improve my writing, and become a more confident biostatistician. I would also like to thank Drs. Roshan Bastani, Beth Glenn, Alison Herrmann, and Fola May and the other researchers at the Center for Cancer Prevention and Control Research for helping me further develop my collaboration and communication skills. Finally, I'd like to thank Drs. Hilary Aralis, Tom Belin, and Beth Glenn for serving on my doctoral committee.

Thank you to my friends and family who have cheered me on throughout this journey. And a huge thank you to my parents and siblings who have done so much for me. Their encouragement and support have been endless.

CURRICULUM VITAE

2024	Ph.D. Candidate Biostatistics, UCLA, Los Angeles, California
2020-2024	Graduate Student Researcher, UCLA Center for Cancer Prevention and Control Research, Los Angeles, California
2020	M.S. Biostatistics, UCLA, Los Angeles, California
2019-2020	Graduate Student Researcher, UCLA Center for Health Policy Research, Los Angeles, California
2017-2019	Research Analyst, Melissa, Rancho Santa Margarita, California
2016	B.S. Probability and Statistics, UCSD, San Diego, California

PUBLICATIONS

Avila, A., Glenn, B. A., Bastani, R., Crespi, C. M. (2024). Handling missing outcome data in cluster randomized trials with both individual- and cluster-level dropout. In preparation.

Crespi, C. M., Gao, S., Payne, A., Nobari, T. Z., Avila, A., Nau, C., . . . & Wang, M. C. (2021). Longitudinal trajectories of adiposity-related measures from age 2–5 years in a population of low-income Hispanic children. *Pediatric Research*, 89(6), 1557-1564.

CHAPTER 1

Introduction

Cluster randomized trials (CRTs) are trials in which clusters of individuals are randomized to different treatment arms. The clusters are naturally occurring or self-selected groups such as schools, hospitals, or geographical areas, and outcomes are observed on the students, patients, or residents within the clusters. CRTs can be used to compare interventions delivered at the cluster level or when individual-level randomization carries a risk of contamination between the intervention and control conditions. The CRT design is a practical choice for many interventions, such as new education programs in schools or workflow changes in hospitals. The use of CRT designs has increased rapidly since the 1980's (Moberg and Kramer, 2015). The NIH (2022) has reported that every 5 years from 1995 to 2015, the number of PubMed abstracts identifying the use of CRTs has doubled.

Missing outcome data in CRTs are common. Missing outcome data in CRTs can occur due to individual dropout, when some individuals within a cluster provide outcome data but others do not. Alternatively, missing outcome data can occur due to cluster dropout, when an entire cluster is lost to follow-up, resulting in missing outcomes for all of the individuals in the cluster. Following other authors' terminology, we refer to the former type of missingness as *sporadically* missing data and the latter as *systematically* missing data (Audigier et al., 2018; Jolani, 2018; Resche-Rigon and White, 2018). While systematically missing data are less common, both types of missingness can be observed in the same study (Isensee et al., 2012; Acosta et al., 2019). A systematic review of the presence and handling of missing data in 86 CRTs from 2013-2014 found that 93% reported sporadically missing outcome data with a median of 19% of individuals missing an outcome (Fiero et al., 2016). Systematically missing data was reported in 27 trials (31%) with a median of 7% of clusters missing all follow-up data.

1.1 Missing Data Methods in CRTs

Although missing outcome data are common in CRTs, it is seldom handled in an optimal manner. A systematic review by Fiero et al. (2016) found that among the 80 CRTs with missing outcome data, complete case analysis, also known as listwise or casewise deletion, was the most common method for handling the missing data ($n = 44$, 55%). While easy to implement, complete case analysis has well-known disadvantages including reduced sample size and power and the possibility of biased estimates (Little and Rubin, 2002; van Buuren, 2018; Enders, 2022). Small percentages of the trials used single or multiple imputation (8% and 2%, respectively). None of the studies used a multiple imputation method designed for multilevel data. A sensitivity analysis of the assumed missing data mechanism was only performed for 14 trials (16%).

Multiple imputation (MI) is an alternative to complete case analysis that avoids reduced sample size and power while providing inferences that reflect variability and uncertainty due to the missing data, and allows for standard complete-data analysis methods to be used (Little and Rubin, 2002; Harel and Zhou, 2007; Li et al., 2015). Inverse probability weighting which is another popular method for handling missing data; however, multiple imputation is often more flexible and efficient (Rubin, 1996; Carpenter and Smuk, 2021; Little et al., 2022). Two of the most common MI approaches are joint modeling (JM) and fully conditional specification (FCS). The standard JM and FCS approaches were created for single-level data and are generally not appropriate for multilevel data. Multilevel MI methods have been developed to handle sporadically missing data (Schafer and Yucel, 2002; van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2011) and data with both sporadically and systematically missing values (Jolani, 2018; Resche-Rigon and White, 2018). However, these methods are not commonly used in practice and applications to CRTs are lacking. Huque et al. (2020) compared the performance of multilevel MI methods within the context of CRTs with only sporadically missing outcome data. Audigier et al. (2018) compared the performance of multilevel MI methods on covariates with sporadically and systematically missing values in the context of meta-analysis. To our knowledge, the performance of multilevel MI methods has not been evaluated for CRTs where outcome data are sporadically and systematically missing due to individual and cluster dropout.

MI methods enable inference under a missing at random (MAR) assumption that the missing values may depend on observed data but not on missing values (Little and Rubin, 2002). It is often prudent to conduct sensitivity analyses to evaluate the robustness of inferences under various missing not at random (MNAR) assumptions that allow missing values to depend on unobserved data. When a CRT has both individual and cluster dropout, the mechanisms leading to the two types of dropout may be different. Hence it would be useful to have sensitivity analysis approaches that allow for different missing data mechanisms for individual and cluster dropout.

1.2 Motivating Examples

In this section we introduce two motivating examples. In Section 1.2.1 we discuss the ABC Healthy Me study which first motivated the work presented in this dissertation. The outcome data in this CRT were both systematically and sporadically missing, and the underlying missing data mechanism was potentially different for the two types of dropout. We then introduce the Korean Healthy Life Project study in Section 1.2.2. This CRT had a similar study design and patterns of missingness as the ABC Healthy Me study, and we use it as the real data application in Chapter 4.

1.2.1 ABC Healthy Me

The University of California, Los Angeles (UCLA) partnered with the Child Care Resource Center (CCRC) to conduct the ABC Healthy Me study. This study is an NIH-funded study (NIH/National Institute of Child Health and Human Development R01HD091136) (Bastani, 2018). CCRC is a large, non-profit organization that works to ensure comprehensive preschool experiences, particularly in low resource and diverse communities. The ABC Healthy Me study is a CRT to test an intervention to decrease the prevalence of obesity among preschool-aged children, implemented with a sample of largely Hispanic, preschool students in Los Angeles County. The trial has a two-group design with cluster randomization at the level of the preschool. The study includes preschools recruited in 4 cohorts over 4 instructional years between 2019 and 2023. Preschool students ages 24-60 months (baseline collection falling on or before the child's 5th birthday) were eligible for participation. Data analysis for this study is ongoing.

Parents in recruited preschools were invited to participate in the ABC Healthy Me study. For those who opted in to have data collected for their child and/or for themselves, baseline data collection occurred at the beginning of the school year and included surveys (parents) and height and weight measurements (children). Following baseline collection, intervention sites participated in the intervention which included implementation of new school-wide policies, classroom curricula, and parent education programs related to nutrition and physical activity. Follow-up data collection was planned to occur at the end of the 10-month instructional year and mirror baseline data collection. The primary outcome is child BMI z-score. BMI z-scores are calculated from children's age, sex, height, and weight. A BMI z-score is the number of standard deviations a child's BMI score is away from the mean (for their age and sex) based on a reference population. BMI z-score was selected as an adiposity index due to its extensive use in childhood obesity research (Braun et al., 2018; Mendoza et al., 2014; Sadeghi et al., 2019).

There were 48 preschools sites (25 intervention and 23 control) included in the study across the 4 cohorts. The first and last cohorts occurred over the 2019-2020 and 2022-2023 school years, respectively. Parent consent was given and data were collected for 1136 eligible students. The number of student participants per preschool ranged from 5-76 with an average of 24 per site. We observed 1019 children at baseline and 837 children at follow-up. There were 117 children who entered the study at follow-up and did not have baseline data.

Looking specifically at children who had baseline data collected but were missing follow-up measurements, such dropout occurred at both the preschool and student level. At the beginning of the COVID-19 pandemic, 3 control sites in cohort 1 dropped out of the study. Baseline data were collected for these sites, but no follow-up data were collected. One cohort 2 control site also dropped out of the study after baseline data collection. Sixty of the 1019 children (6%) who had baseline height and weight data had systematically missing outcomes due to these 4 sites dropping out. The remaining dropout (239 children, 23%) was sporadic, occurring at the individual level with children leaving schools due to the pandemic (in cohorts 1 and 2) or for other reasons or being absent when measurements were collected. Individual dropout and cluster dropout may be due to different mechanisms. While systematic dropout was solely due to the COVID-19 pandemic, some cases of sporadic dropout occurred for other reasons. Ultimately, 29% of the children for which we collected baseline data were missing follow-up data. Table 1.1 gives the frequency of each type of missingness (systematic or sporadic) among the sample of patients who had baseline data collected. Students missing baseline data are not included in this table.

In addition to the children with missing follow-up measurements, we also observed 117 children (69 intervention and 48 control) with sporadically missing baseline data. This occurred when children were absent at the time of baseline data collection or enrolled in the school after baseline data collection. These children may have had less exposure to the intervention. Age and sex values were also missing from the baseline data for 18 children.

Table 1.1
Outcome missingness by treatment arm and overall for students with baseline data

	Total n		Missing n (%)		
Arm	Clusters	Individuals	Clusters	Individuals	
				Systematic	Sporadic
Intervention	25	549	0 (0.0)	0 (0.0)	127 (23.1)
Control	23	470	4 (17.4)	60 (12.8)	112 (23.8)
Overall	48	1019	4 (8.3)	60 (5.9)	239 (23.5)

These variables are needed to calculate BMI z-score.

The planned analysis model for the ABC Healthy Me data is a linear mixed-effects model (LMM) with a random intercept for school, but data analysis is still ongoing. Therefore, we identified a different CRT conducted by UCLA, the Korean Healthy Life Project, that has a similar study design and patterns of missingness as the ABC Healthy Me study data and has already been published.

1.2.2 Korean Healthy Life Project

The motivating example that we use as the focus of our real data application is from the Korean Healthy Life Project (KHLP) which was supported by NIH grant P01 CA109091 (Bastani et al., 2015). The KHLP was a two-arm CRT to evaluate a church-based intervention to improve hepatitis B virus (HBV) knowledge and testing among Korean-Americans in the Los Angeles area. The study was conducted between 2006 and 2012. In this study, 52 churches with Korean-American congregations were randomized to an intervention or a control condition. The number of participants from each church ranged from 7-71 with an average of 22 per church. Participants at churches assigned to the intervention attended a group session on liver cancer and HBV testing, and participants at control churches attended a session on physical activity and nutrition. Participants completed self-report surveys at baseline and 6-month follow-up. The study had several outcomes of interest. For the purposes of this dissertation, we focus on HBV knowledge score, which was measured via a 9-item module and yielded a possible score of 0 to 9.

Although none of the churches dropped out of the study entirely, three churches, two in the intervention arm and one in the control arm, hosted HBV-related events led by outside organizations after baseline data collection. The outcomes of individuals from these churches were thus considered to be contaminated and the investigators considered it advisable to conduct analyses regarding these outcomes as missing. For our data application, we consider these churches to have systematically missing outcomes. These three sites had a total of 130 participants, which was about 12% of the total of 1,123 participants. An additional 148 participants (13% of the total) had sporadically missing outcomes. Overall, 25% of the participants had missing follow-up knowledge scores when outcomes at the three sites are dropped. Table 1.2 summarizes missingness by treatment arm.

Table 1.2
Outcome missingness by treatment arm and overall

	Total n		Missing n (%)		
Arm	Clusters	Individuals	Clusters	Individuals Systematic	 Sporadic
Intervention	26	543	2 (7.7)	103 (19.0)	75 (13.8)
Control	26	580	1 (3.8)	27 (4.7)	73 (12.6)
Overall	52	1123	3 (5.8)	130 (11.6)	148 (13.2)

In this study, individual dropout and cluster “dropout” due to contamination are clearly due to different mechanisms. Hence in a sensitivity analysis evaluating the robustness of inferences to assumptions about the missing values, it would be helpful to explicitly allow for different assumptions about these two types of missing values.

This study also featured missing baseline knowledge scores as part of the study design. Churches were randomized to receive either a short or long form of the survey at baseline. The short form was given to 24 (46%) of the churches, 12 in each treatment arm, and did not include the 9-item knowledge score module. This resulted in systematically missing baseline knowledge scores for 568 (51%) participants that were missing completely at random (MCAR). While the main focus of this dissertation is handling missing outcome data, we also consider how a missing baseline covariate affects the performance of the multiple imputation methods.

To understand factors associated with missing outcomes, we compared the baseline demographic characteristics among completers and those who only had baseline data. Although cluster-level non-response is due to contamination, we refer to both cluster-level and individual-level non-responders as "dropouts". Dropouts with systematically missing outcomes were more likely to be in the intervention arm, attend large churches, attend a church in Koreatown, and be unmarried compared to completers. Dropouts with sporadic missingness were more likely to attend small churches, have never been married, have been born outside Korea, and were younger compared to completers (Table 1.3).

Bastani et al. (2015) found no statistically significant baseline demographic differences between intervention and control participants. The distribution of knowledge score among completers at baseline and follow-up by treatment arm is reported in Table 1.4. Mean knowledge score at baseline was similar between groups. At follow-up, intervention participants had a higher mean knowledge score by 0.5 units compared to control participants.

The KHLP study highlights the challenges created by missing data in a CRT. The target analysis model is a LMM with a random intercept. Significant dropout at both the individual level and cluster level make analyzing the data and obtaining reliable results challenging.

In this dissertation, we develop a well-performing and practical approach to handle inference in CRTs when outcome data may be both sporadically and systematically missing. To do so, we first compare the performance of available multilevel MI methods in the context of CRTs with both sporadically and systematically missing outcome data. We then develop methods for conducting sensitivity analyses based on the best-performing method. The

Table 1.3
Baseline demographics overall and by completion status

Characteristics	Completer	Dropout		Total	p-value	
	n (%)	n (%)		n (%)		
		Systematic	Sporadic		Systematic	Sporadic
Treatment arm						
Control	480 (56.8)	27 (20.8)	73 (49.3)	580 (51.6)	<0.001	0.09
Intervention	365 (43.2)	103 (79.2)	75 (50.7)	543 (48.4)		
Size						
51-200	351 (41.5)	0 (0.0)	73 (49.3)	424 (37.8)	<0.001	0.02
201-900	297 (35.1)	0 (0.0)	55 (37.2)	352 (31.3)		
900+	197 (23.3)	130 (100.0)	20 (13.5)	347 (30.9)		
Location						
Koreatown	273 (32.3)	130 (100.0)	57 (38.5)	460 (41.0)	<0.001	0.14
Non-Koreatown	572 (67.7)	0 (0.0)	91 (61.5)	663 (59.0)		
Age						
Mean (sd)	46.3 (11.9)	45.4 (12.8)	42.1 (14.0)	45.6 (12.4)	0.58	0.001
Sex						
Female	556 (65.8)	88 (67.7)	89 (60.1)	733 (65.3)	0.67	0.18
Male	289 (34.2)	42 (32.3)	59 (39.9)	390 (34.7)		
Marital Status						
Married	661 (78.3)	86 (66.2)	93 (63.3)	840 (74.9)	0.01	<0.001
Single	79 (9.4)	19 (14.6)	10 (6.8)	108 (9.6)		
Never married	104 (12.3)	25 (19.2)	44 (29.9)	173 (15.4)		
Missing	1	0	1	2		
College graduate						
No	387 (46.0)	60 (46.2)	73 (49.7)	520 (46.5)	0.97	0.41
Yes	455 (54.0)	70 (53.8)	74 (50.3)	599 (53.5)		
Missing	3	0	1	4		
English fluency						
Fluent	178 (21.1)	23 (17.7)	33 (22.4)	234 (20.9)	0.38	0.71
Not fluent	667 (78.9)	107 (82.3)	114 (77.6)	888 (79.1)		
Missing	0	0	1	1		
Country of birth						
Korea	821 (97.2)	128 (98.5)	138 (93.2)	1087 (96.8)	0.56	0.02
US/Other	24 (2.8)	2 (1.5)	10 (6.8)	36 (3.2)		
Income						
<$30k	162 (23.3)	30 (29.1)	30 (28.8)	222 (24.6)	0.54	0.46
$30k - $50k	196 (28.2)	28 (27.2)	25 (24.0)	249 (27.6)		
$50k - $-80k	169 (24.3)	25 (24.3)	28 (26.9)	222 (24.6)		
>$80k	168 (24.2)	20 (19.4)	21 (20.2)	209 (23.2)		
Missing	150	27	44	221		
Baseline knowledge						
Mean (sd)	6.1 (1.5)	5.9 (1.8)	5.9 (1.8)	6.0 (1.6)	0.80	0.70
Missing	408	85	76	569		

p-values based on Wilcoxon rank sum and chi-square tests comparing indicated dropouts to completers

Table 1.4
Knowledge score distribution at baseline and follow-up by treatment arm

	Control	Intervention	Total
Knowledge score	mean (sd)	mean (sd)	mean (sd)
Baseline	6.1 (1.5)	6.0 (1.6)	6.0 (1.6)
Missing	310	259	569
Follow-up	6.3 (1.5)	6.8 (1.3)	6.5 (1.5)
Missing	100	178	278

sensitivity analysis approach is user-friendly and tests the robustness of inferences under different missing not at random (MNAR) assumptions regarding the missing data, allowing for potentially different MNAR assumptions for sporadically and systematically missing data. We restrict attention to continuous outcomes.

The rest of this dissertation is organized as follows. In Chapter 2 we introduce notation and describe the multilevel MI methods that we evaluate, and we compare the performance of those MI methods via a simulation study. Chapter 4 discusses the MNAR sensitivity analysis methods we developed for evaluating the robustness of our MAR analysis, and we apply the methods to our motivating example. Finally, in Chapter 5 we conclude with a discussion.

CHAPTER 2

Multiple Imputation Methods

The analysis of our motivating example described in Section 1.2.2 requires identifying a MI method that is appropriate for use when the target analysis is a LMM and that accommodates incomplete outcome data at both the cluster and individual level. In Section 2.1, we define various methods for MI of multilevel data.

Let $\mathbf{Y}_{nxp} = (\mathbf{y}_1, \ldots, \mathbf{y}_p)$ be a matrix containing data for a total of n units on p potentially incomplete variables. Let k (k in 1, $\ldots$, p) denote one of the p variables, i (i in 1, $\ldots$, m) denote the cluster index, and j (j in 1, $\ldots$, n_i) denote the index for individuals in cluster i. Then $\mathbf{y}_k$ denotes the k^{th} variable and $\mathbf{y}_{ki}$ denotes the vector of variable values $\mathbf{y}_k$ restricted to individuals within cluster i. Let $(\mathbf{y}_k^{obs}, \mathbf{y}_k^{mis})$ be the observed and missing parts of $\mathbf{y}_k$ and let $\mathbf{Y}^{obs} = (\mathbf{y}_1^{obs}, \ldots, \mathbf{y}_p^{obs})$ and $\mathbf{Y}^{mis} = (\mathbf{y}_1^{mis}, \ldots, \mathbf{y}_p^{mis})$. The response indicator matrix $\mathbf{R}$ is a nxp matrix with elements r_{kij}. If y_{kij} is observed, then $r_{kij} = 1$, and if y_{kij} is missing, then $r_{kij} = 0$. If we define $\boldsymbol{\psi}$ as the parameters of the missing data model, then the missing data model can be expressed as $\mathrm{P}(\mathbf{R} \,|\, \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \boldsymbol{\psi})$.

Using this notation, the three missing data mechanism classes, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), can be defined as follows (Rubin, 1976). The data are said to be MCAR if $\mathrm{P}(\mathbf{R} = \mathbf{0} \,|\, \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \boldsymbol{\psi}) = \mathrm{P}(\mathbf{R} = \mathbf{0} \,|\, \boldsymbol{\psi})$. If $\mathrm{P}(\mathbf{R} = \mathbf{0} \,|\, \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \boldsymbol{\psi}) = \mathrm{P}(\mathbf{R} = \mathbf{0} \,|\, \mathbf{Y}^{obs}, \boldsymbol{\psi})$, the data are MAR. Lastly, the data are said to be MNAR if $\mathrm{P}(\mathbf{R} = \mathbf{0} \,|\, \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \boldsymbol{\psi})$ does not simplify, meaning the probability to be missing depends on the missing data itself.

2.1 MI Methods for Linear Mixed Models

Our analysis model of interest is a linear mixed-effects model (LMM), a common model for clustered data (Laird and Ware, 1982). Given our target analysis model, multilevel MI methods using a LMM as the imputation model are of specific interest. Let $\mathbf{y}_{ki}$ be an incomplete n_i x 1 vector of continuous outcomes for the n_i individuals in cluster i. The imputation model with parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{G}, \mathbf{D}_i)$ is

$$\begin{aligned} &\mathbf{y}_{ki} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \\ &\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \mathbf{G}), \\ &\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{D}_i) \end{aligned} \tag{2.1}$$

where $\mathbf{X}_i$ $(n_i \times q)$ and $\mathbf{Z}_i$ $(n_i \times q')$ are the design matrices for fixed effects and random effects, respectively, $\boldsymbol{\beta}$ is the q-vector of regression coefficients of fixed effects, $\boldsymbol{\gamma}_i$ is the q'-vector of random effects for cluster i, $\mathbf{G}$ $(q' \times q')$ is the between cluster covariance matrix, and $\mathbf{D}_i = \sigma_i^2\mathbf{I}(n_i)$ $(n_i \times n_i)$ is the covariance matrix for observations within cluster i. It is often assumed that the residual variance is equal for all clusters: $\sigma_i^2 = \sigma^2$ (homoscedasticity).

Several extensions of the standard JM and FCS approaches to deal with multilevel data have been proposed (Huque et al., 2020). Shafer and Yucel proposed a joint multivariate linear mixed-effects model approach (JM-MLMM) (Schafer and Yucel, 2002). This approach uses a multivariate LMM to impute all incomplete variables. A FCS adaption of the JM-MLMM approach (FCS-LMM) has also been developed (van Buuren and Groothuis-Oudshoorn, 2011). The FCS-LMM method uses a LMM for imputing missing values in each incomplete variable given all the others. When there is more than one incomplete variable, rather than using a multivariate LMM, the method cycles iteratively through the univariate LMM imputation models (one for each incomplete variable). Both of these methods assume there is a constant residual variance across all clusters. van Buuren extended the FCS-LMM approach to allow for heteroscedastic (cluster-specific) residual variances (van Buuren, 2011). We denote this approach as FCS-LMM-het.

The performance of the JM-MLMM, FCS-LMM, and FCS-LMM-het methods were compared by Huque et al. (2020). The methods were used in the context of clustered data where the outcome and a covariate are incomplete. When the analysis was a LMM with a random intercept and a random slope associated with the incomplete covariate, the authors found that FCS-LMM and FCS-LMM-het performed best in the estimation of regression parameters and variance components. The JM-MLMM method was incompatible with a LMM with random intercepts and slopes if both the outcome and random-slope covariate were incomplete.

For our purposes, the analysis model of interest is a LMM with a random intercept for cluster. Because we are interested in data with a similar structure, pattern of missingness, and analysis model as the data setting considered by Huque et al. (2020), the JM-MLMM, FCS-LMM, and FCS-LMM-het models are good candidate methods to evaluate for our purposes. However, since Huque et al. found that the JM-MLMM method performed similarly or worse than the FCS methods, we chose to focus on FCS methods for this dissertation. The FCS methods also allow for more flexibility when choosing imputation models for variables (Carpenter and Smuk, 2021; Little et al., 2022). Additionally, there are FCS method extensions developed to deal with both sporadically and systematically missing data.

In the following sections, we describe each MI method that we considered in our simulation study. A single-level imputation method was also included for the sake of comparison.

2.1.1 Single-Level MI Methods

Standard JM and standard FCS are designed for single-level data. The single-level methods have been applied to impute multilevel data by either ignoring the clusters or treating them as fixed effects. It is known that the estimation of the intraclass correlation (ICC) is affected by both approaches (van Buuren, 2018). When ignoring clusters, the ICC is underestimated. When treating clusters as fixed effects by using dummy variables, the ICC is overestimated. The use of $m - 1$ dummy variables for m clusters can also reduce degrees of freedom. In both cases, parameter estimates are often biased.

Although single-level MI methods are not expected to perform well in the context of multilevel data, given that they are more familiar and widely utilized than multilevel MI methods, there is interest in seeing how a single-level method performs compared to methods that account for clustering. Therefore, we decided to include a standard FCS (FCS-stnd) method that ignores clustering in our simulation study. This method was implemented using the *mice.impute.norm* function of the *mice* R package (van Buuren and Groothuis-Oudshoorn, 2011).

2.1.2 FCS-LMM

The FCS-LMM method, developed by van Buuren and Groothuis-Oudshoorn (2011), cycles iteratively through the univariate LMM imputation models for each incomplete variable. First, distributions of the parameters are simulated by Markov chain Monte Carlo (MCMC) methods. Then, the imputed value is drawn from the conditional distribution of the missing observations given the already drawn parameter values.

Let $\mathbf{y}_k$ be a continuous incomplete variable vector. Under the MAR assumption, imputations can be generated as follows:

1. Sample $\boldsymbol{\beta}^*$ from $P(\boldsymbol{\beta} \mid \mathbf{y}_k^{obs}, \boldsymbol{\gamma}, \sigma^2)$
2. Sample $\boldsymbol{\gamma}_i^*$ from $P(\boldsymbol{\gamma} \mid \mathbf{y}_k^{obs}, \boldsymbol{\beta}^*, \mathbf{G}, \sigma^2)$
3. Sample $\mathbf{G}^*$ from $P(\mathbf{G} \mid \mathbf{y}_k^{obs}, \boldsymbol{\gamma}^*)$
4. Sample σ^{*2} from $P(\sigma^2 \mid \mathbf{y}_k^{obs}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$

5. Repeat step 1-4 until convergence

6. Sample $\mathbf{y}_k^{mis*}$ from $p(\mathbf{y}_k^{mis} \mid \mathbf{y}_k^{obs}, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{G}^*, \sigma^{*2})$.

Under the model in Equation 2.1, the imputation for $\mathbf{y}_{ki}$, $\mathbf{y}_{ki}^*$, is calculated by drawing

$$\begin{aligned} &\mathbf{e}_i^* \sim N(0, \sigma^{*2}\mathbf{I}(n_i)), \\ &\mathbf{y}_{ki}^* = \mathbf{X}_i\boldsymbol{\beta}^* + \mathbf{Z}_i\boldsymbol{\gamma}_i^* + \mathbf{e}_i^* \end{aligned} \tag{2.2}$$

where all parameters on the right-hand side of the equations are the values drawn from the Gibbs sampler. This method was implemented using the *mice.impute.2lpan* function of the *mice* package in R (van Buuren and Groothuis-Oudshoorn, 2011).

2.1.3 FCS-LMM-het

The FCS-LMM method above produces imputations under an LMM model that assumes all clusters have the same within-cluster variance σ^2. The FCS-LMM-het method, introduced by van Buuren (2011), produces imputations under a more general LMM model in which the within cluster variance σ_i^2 is allowed to vary over clusters. The Gibbs sampler for this heterogeneous model replaces step 4 above with

4. Sample σ_i^{*2} from $P(\sigma_i^2 \,|\, \sigma_0^2, \phi) \sim \frac{\sigma_0^2\chi_1^2}{\phi}$

where σ_0^2 and ϕ are hyperparameters specifying the location of prior belief about residual variance σ_i^2 and a measure of variability of the variances σ_i^2, respectively (Kasim and Raudenbush, 1998). The FCS-LMM-het method was developed as a more general version of the FCS-LMM method which did not always produce good imputations for incomplete predictors (van Buuren, 2018). This method was implemented using the *mice.impute.2l.norm* function of the *mice* package in R (van Buuren and Groothuis-Oudshoorn, 2011).

2.2 FCS Methods for Systematically Missing Data

The MI methods discussed above were developed to deal with data that are sporadically missing values only. Methods that handle both sporadically and systematically missing data have been developed by several authors. Snijders and Bosker (2011) proposed using the maximum likelihood estimates (MLE) obtained from the univariate LMM to draw missing values and Robitzsch and Grund (2021) extended the method to handle systematically missing values. We refer to this method as FCS-LMM-MLE. The method developed by Jolani (2018) uses a Bayesian formulation of the univariate LMM to draw missing values. We use the acronym FCS-GLM (GLM for "generalized linear model") to reference this method. The method developed by Resche-Rigon and White (2018) fits a two-stage estimator using only clusters with sporadically missing data and then approximates the posterior distribution. We refer to this method as FCS-2stage.

Generally, the imputation step in MI uses the predictive distribution $P(\mathbf{Y}^{mis}|\mathbf{Y}^{obs})$ to obtain missing values. If we specify an imputation model with parameter $\boldsymbol{\theta} = (\boldsymbol{\beta},\ \mathbf{G},\ \sigma^2)$, then missing values are obtained by:

1. Fitting the imputation model (2.1) to the observed data to obtain an estimate (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{G}}, \hat{\sigma}^2)$ and $\hat{\boldsymbol{\gamma}}_i$.

2. Drawing $\boldsymbol{\theta}^*$ from $P(\boldsymbol{\theta}|\mathbf{Y}^{obs})$, its posterior distribution.

3. Drawing missing data according to $P(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \boldsymbol{\theta}^*)$.

The FCS-LMM-MLE, FCS-GLM, and FCS-2stage methods use different approaches for each of these steps. All methods modify step 3 of this process to accommodate systematically missing values. The value $\boldsymbol{\gamma}_i^*$ is drawn from a different distribution for clusters where $\mathbf{y}_{ki}$ is systematically missing versus sporadically missing.

2.2.1 FCS-LMM-MLE

The FCS-LMM-MLE method uses the maximum likelihood estimates (MLEs) obtained from fitting a LMM to observed data in order to multiply impute missing values. The imputation process can be broken into three steps with the third step designed to accommodate systematically missing values. If we assume the error terms are homoscedastic, then the imputation is performed as follows:

1. Fit the imputation model (2.1) to the observed data to obtain an estimate (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{G}}, \hat{\sigma}^2)$ and $\hat{\boldsymbol{\gamma}}_i$.

2. Estimate $\boldsymbol{\theta}^*$ by:

    - Drawing $\boldsymbol{\beta^*}$ from $N(\hat{\boldsymbol{\beta}}, var(\hat{\boldsymbol{\beta}}))$
    - Setting $\mathbf{G}^*$ equal to the posterior variance of $\boldsymbol{\gamma}_i$
    - Setting $\sigma^{*2} = \hat{\sigma}^2$.

3. Simulate $P(\mathbf{y}_k^{mis}|\mathbf{Y}^{obs}, \boldsymbol{\theta}^*)$ by drawing:

    - $\boldsymbol{\gamma}_i^*$ from $N(0, \hat{\mathbf{G}})$ for all clusters where $\mathbf{y}_{ki}$ is systematically missing
    - $\boldsymbol{\gamma}_i^*$ from $N(\hat{\boldsymbol{\gamma}}_i, \mathbf{G}^*)$ for all clusters where $\mathbf{y}_{ki}$ is sporadically missing
    - $\mathbf{y}_{ki}^{mis}$ from $N(\mathbf{X}_i\boldsymbol{\beta}^* + \mathbf{Z}_i\boldsymbol{\gamma}_i^*, \sigma^{*2}\mathbf{I}(n_i))$.

The FCS-LMM-MLE method has been extended to binary variables but not to include the heteroscedasticity assumption (Robitzsch and Grund, 2021). This method was implemented using the *mice.impute.2l.continuous* function of the *miceadds* package in R (Robitzsch and Grund, 2021).

2.2.2 FCS-GLM

The FCS-GLM method uses an approximate Bayesian approach to multiply impute data. Similar to the FCS-LMM-MLE method, a three step imputation process with the third step designed to accommodate systematically missing values is used. Assuming the error terms are homoscedastic, the imputation is performed as follows:

1. Fit the imputation model (2.1) to the observed data to obtain an estimate (MLE) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{G}}, \hat{\sigma}^2)$ and $\hat{\boldsymbol{\gamma}}_i$.

2. Draw $\boldsymbol{\theta}^*$ from $P(\boldsymbol{\theta}|\mathbf{Y}^{obs})$, where the posterior distributions of the parameters are derived using an approximate Bayesian formulation of the LMM based on non-informative priors.

3. Simulate $P(\mathbf{y}_k^{mis}|\mathbf{Y}^{obs}, \boldsymbol{\theta}^*)$ by drawing:

   - $\boldsymbol{\gamma}_i^*$ from $N(0, \mathbf{G}^*)$ for all clusters where $\mathbf{y}_{ki}$ is systematically missing
   - $\boldsymbol{\gamma}_i^*$ from $P(\boldsymbol{\gamma}_i|\mathbf{y}_{ki}^{obs}, \boldsymbol{\theta}^*)$ for all clusters where $\mathbf{y}_{ki}$ is sporadically missing
   - $\mathbf{y}_{ki}^{mis}$ from $N(\mathbf{X}_i\boldsymbol{\beta}^* + \mathbf{Z}_i\boldsymbol{\gamma}_i^*, \sigma^{*2}\mathbf{I}(n_i))$.

Exact specification of the posterior distributions and distributions from which $\boldsymbol{\gamma}_i^*$ is drawn can be found in the Jolani (2018) paper. The method has been extended to binary variables but not to include the heteroscedasticity assumption (Audigier et al., 2018). The authors present the FCS-GLM method as a way to impute incomplete predictors. We examine how this method performs for an outcome variable with both sporadically and systematically missing data. The FCS-GLM method can be implemented using the *mice.impute.2l.glm.norm* function of the *micemd* package in R (Audigier and Resche-Rigon, 2021).

2.2.3 FCS-2stage

The FCS-2stage method uses a FCS approach based on a two-stage estimator. Under the case $\mathbf{X}_i = \mathbf{Z}_i$, the imputation model (2.1) is rewritten as

$$\begin{aligned} \mathbf{y}_{ki} &= \mathbf{X}_i(\boldsymbol{\beta} + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\gamma}_i &\sim N(\mathbf{0}, \mathbf{G}), \\ \boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \sigma_i^2 \mathbf{I}(n_i)). \end{aligned} \tag{2.3}$$

Let $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\gamma}_i$. At stage one, the two-stage estimator $\hat{\boldsymbol{\beta}}_i$ is fit by computing the maximum likelihood estimator of a linear model separately within each cluster without systematically missing data:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}_{ki}. \tag{2.4}$$

At stage two, the results are combined using a multivariate random-effects meta-analysis model:

$$\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta} + \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}'_i \tag{2.5}$$

where $\boldsymbol{\gamma}_i \sim N(0, \mathbf{G})$ and $\boldsymbol{\epsilon}'_i \sim N(0, \sigma_i^2 (\mathbf{X}_i^T \mathbf{X}_i)^{-1})$. The parameters $\boldsymbol{\beta}$ and $\mathbf{G}$ can be estimated by restricted maximum likelihood or by method of moments, which can be faster (Resche-Rigon and White, 2018; Audigier et al., 2018). The imputation of $\mathbf{y}_k$ is then performed as follows:

2. Draw $\boldsymbol{\theta}^*$ according to the asymptotic posterior.

3. Generate $P(\mathbf{y}_k^{mis} | \mathbf{Y}^{obs}, \boldsymbol{\theta}^*)$ by drawing:

    - $\boldsymbol{\gamma}_i^*$ from $N(0, \mathbf{G}^*)$ for all clusters where $\mathbf{y}_{ki}$ is systematically missing

    - $\boldsymbol{\gamma}_i^*$ conditionally on $\hat{\boldsymbol{\beta}}_i$ for all clusters where $\mathbf{y}_{ki}$ is sporadically missing

- y_{kij}^{mis} from $N(\mathbf{x}_{ij}(\boldsymbol{\beta}^* + \boldsymbol{\gamma}_i^*), \sigma_i^{*2})$ for all j such that y_{kij}, the observation j in cluster i, is missing.

Details on the asymptotic posterior distributions for each parameter and the distributions from which $\boldsymbol{\gamma}_i^*$ is drawn can be found in Resche-Rigon and White (2018) and Audigier et al. (2018). The FCS-2stage method allows for heteroscedastic error terms and can be extended to binary variables (Audigier et al., 2018). Unlike the FCS-GLM method, the FCS-2stage method is presented as a way to impute incomplete predictor or outcome variables. The method was implemented using the *mice.impute.2l.2stage.norm* function of the *micemd* package in R (Audigier and Resche-Rigon, 2021).

CHAPTER 3

Simulation Study

A simulation study was conducted with the goal of the identification of multiple imputation (MI) methods that work well for cluster randomized trials with both systematically and sporadically missing outcome data. Performance of the MI methods described in Section 2.1 were compared for the various scenarios. In Section 3.1 we explain the simulation study design, and in Section 3.2 the performance metrics for each MI method are reported.

3.1 Design of Simulation Study

We conducted a simulation study to evaluate the performance of the MI methods described in Section 2.1 on data from a CRT with systematically and sporadically missing outcome data, i.e., cluster and individual dropout. The simulated data were based on the KHLP study. Each simulated data set consisted of 60 clusters (30 control and 30 intervention) with 20 individuals per cluster. Each data set had 6 variables: subject identification (ID) number, cluster ID number, treatment assignment (control or intervention), age, baseline knowledge score, and follow-up knowledge score. Individuals indexed by j are nested within clusters indexed by i. Values of age, treatment assignment (T_i), and baseline knowledge score (knw0_{ij}) were simulated based on their observed distributions in the KHLP data. Follow-up knowledge scores (knw6_{ij}) were generated using the model

$$\text{knw6}_{ij} = (0.45 + \gamma_i) + 0.5T_i + 0.3knw0_{ij} + \epsilon_{ij} \tag{3.1}$$

where $\gamma_i \sim N(0, 0.04)$ is the cluster-specific intercept and $\epsilon_{ij} \sim N(0, 0.96)$.

Missingness was introduced into the outcome under 4 scenarios: (i) sporadically and systematically missing outcome values were both MCAR, (ii) sporadically and systematically missing outcome values were both MAR, (iii) sporadically missing were MCAR and systematically missing were MAR, and (iv) sporadically missing were MAR and systematically missing were MCAR. When values were systematically MAR (i.e., cluster dropout), missingness was dependent on treatment assignment. When values were sporadically MAR (i.e., individual dropout), missingness was dependent on age. In all scenarios, 10% of the outcome data were systematically missing and an additional 20% were sporadically missing. As described in Section 1.2.2, the KHLP study also had systematically missing baseline knowledge scores that were MCAR by design. Therefore, we considered 4 additional scenarios that featured baseline knowledge score systematically MCAR (from 23% of churches in each arm) to the above scenarios.

In the imputation and analysis models, follow-up knowledge scores were regressed on treatment assignment and baseline knowledge score. The models also included a random intercept for church. Under the sporadically MAR scenarios, age was also included as a covariate in the imputation model.

We evaluated the 6 MI methods described in Section 2.1. Three of the multilevel MI methods we considered allow for imputation of sporadically and systematically missing data: FCS-LMM-MLE [*mice.impute.2l.continuous*], FCS-GLM [*mice.impute.2l.glm.norm*], and FCS-2stage [*mice.impute.2l.2stage.norm*]. We also examined FCS-LMM [*mice.impute.2l.pan*] and FCS-LMM-het [*mice.impute.2l.norm*], which were designed for sporadically missing data; however they can be used to produce imputations for entire clusters. For comparison, we also performed full data analysis (analysis of the original data set before missing values were introduced), complete case (CC) analysis (analysis after listwise deletion of the individuals with incomplete data) and MI with FCS-stnd [*mice.impute.norm*], a common single-level MI method that does not account for multilevel structure.

We generated 200 complete data sets. We then introduced missing values to the outcome variable under scenarios (i) - (iv) and applied the MI methods. For each MI method, we created $M = 5$ imputed data sets for a given incomplete data set. We analyzed these imputed data sets using the same analysis model. The 5 estimates of the model's parameters and standard errors were then pooled according to Rubin's rules (Rubin, 1987).

The primary parameters of interest were the coefficient for treatment assignment β_T, the variance components σ^2_{site} and σ^2_{error}, and the ICC. We computed the ICC as $\sigma^2_{site}/(\sigma^2_{site} + \sigma^2_{error})$. The ICC quantifies the degree to which outcomes within a cluster are correlated and affects the standard error of the treatment effect estimator (Donner and Klar, 2000; Moerbeek and Teerenstra, 2015). The ICC is also an important parameter in CRT sample size calculations. Researchers planning new trials often look to existing literature for ICC estimates. Therefore imputation methods should ideally preserve the true ICC.

The performance of the methods in estimating these parameters was assessed by the average estimate, bias, percent bias, root mean squared error (RMSE), root mean square of the estimated standard error (model SE), and the coverage of the associated confidence interval. Let Q be the true value of a parameter. Then bias is the difference between the expected value of the estimate and the true value, $E(\overline{Q}) - Q$. Percent bias is defined as $100 \times |(E(\overline{Q}) - Q)/Q|$, and 5% is considered to be the reasonable upper limit (van Buuren, 2018). RMSE is defined as $\sqrt{(E(\overline{Q}) - Q)^2}$. Coverage is the proportion of confidence intervals that contain the true value.

To assess the level of uncertainty in these results and to ensure that the use of 200 complete data sets provided a suitable level of precision, we calculated the Monte Carlo standard error (MCSE) of the parameters of interest for each simulated scenario (Koehler et al., 2009; White, 2010). The calculations were performed using the *rsimsum* package in R (Gasparini, 2018).

3.2 Simulation Study Results

In this section we report the simulation study results for the model with an incomplete outcome and the model with both an incomplete outcome and covariate. The MCSEs for the simulations are reported in Appendix A. The MCSEs of all parameters across all scenarios were less than 0.04 suggesting that 200 was a sufficient number of replications to achieve minimal uncertainty.

3.2.1 Incomplete Outcome

Tables 3.1 - 3.4 summarize results under missing data mechanism scenarios (i) - (iv). The performance of each MI method was fairly consistent across the 4 scenarios. Every method except FCS-2stage performed well for estimating β_T; the absolute percent bias was over 15% for the FCS-2stage method and under 5% for all other methods. The FCS-GLM method preformed best with regard to coverage with estimates between 93.5% - 94.5%. Coverage was around 95% for the FCS-LMM-MLE method except under scenario (iv), for which the coverage was 98%. Coverage ranged from 97% to 98% for the FCS-LMM and FCS-LMM-het methods under all scenarios.

When estimating σ_{site}, the absolute percent bias was over 5% for all methods except FCS-LMM-MLE under all scenarios. The single-level FCS-stnd method and the FCS-GLM method underestimated σ_{site} while the other multilevel methods overestimated this parameter. The absolute percent bias was lowest for the FCS-LMM-MLE method, with values under 5% for all scenarios. The FCS-2stage and standard FCS method had particularly high absolute percent bias. This was expected for the standard FCS method because clustering is not accounted for, and suggests the FCS-2stage method may not account for clustering properly in the scenarios we considered. The absolute percent bias for CC analysis was over 5% for all scenarios. All methods performed similarly for estimating σ_{error}, with absolute percent bias under 5%. Across all scenarios, the FCS-LMM-MLE method returned an ICC estimate closest to the true value. The FCS-LMM-het and FCS-2stage methods were the most computationally intensive, taking 1.9-2.2 seconds per data set.

3.2.2 Incomplete Outcome and Covariate

We considered 4 additional simulation scenarios in which both the outcome and a covariate were incomplete. For each scenario, systematic missingness was introduced to baseline knowledge score completely at random, as it was in our motivating example. The results can be found in Tables 3.5 - 3.8. The FCS-LMM-het method did not run when the systematically missing covariate was introduced. The results for the other methods were similar to those observed when only the outcome was missing with one exception. Similar to when only the outcome was missing, the FCS-2stage performed poorly at estimating β_T and σ_{site} across all scenarios. However, the absolute percent bias when estimating σ_{error} also exceeded 5% for scenarios (ii) and (iv).

3.3 Discussion

The simulation study results suggest that the FCS-2stage and FCS-stnd methods do not perform well when applied to CRT data with similar characteristics as the KHLP study data. Both methods consistently yielded an absolute percent bias of over 25% when estimating σ_{site} with FCS-2stage overestimating and FCS-stnd underestimating the parameter. The FCS-2-stage method also consistently underestimated β_T. The bias obtained by the FCS-2stage method is consistent with the findings of Audigier et al. and could be attributed to the small cluster sizes (Audigier et al., 2018). Its performance might improve with larger cluster sizes.

The estimate of σ_{site} was underestimated using CC analysis with an absolute percent bias of over 5% for all scenarios. This may be due to the exclusion of clusters with systematically

Table 3.1
Simulation results under the scenario with sporadically and systematically MCAR outcome data. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
est	0.501	0.502	0.499	0.503	0.504	0.504	0.424	0.504
bias	0.001	0.002	-0.001	0.003	0.004	0.004	-0.076	0.004
% bias	0.207	0.448	-0.174	0.652	0.781	0.776	**-15.15**	0.726
model se	0.076	0.087	0.096	0.095	0.093	0.084	0.103	0.080
95% coverage	96.5	96.0	98.0	97.0	96.0	94.5	95.0	95.0
rmse	0.072	0.078	0.082	0.081	0.081	0.080	0.103	0.081
σ_{site}								
est	0.190	0.189	0.227	0.218	0.207	0.166	0.262	0.140
bias	-0.010	-0.011	0.027	0.018	0.007	-0.034	0.062	-0.060
% bias	-4.831	**-5.341**	**13.27**	**9.199**	3.320	**-17.04**	**31.19**	**-30.12**
σ_{error}								
est	0.981	0.982	0.981	0.983	0.993	0.983	0.982	0.991
bias	0.001	0.002	0.001	0.004	0.013	0.003	0.002	0.011
% bias	0.079	0.194	0.130	0.369	1.356	0.347	0.175	1.117
site ICC	0.038	0.038	0.052	0.049	0.044	0.030	0.068	0.021
average time to MI one data set (sec)			0.16	2.07	0.15	0.58	1.92	0.02

Table 3.2
Simulation results under the scenario with sporadically and systematically MAR outcome data. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
est	0.501	0.504	0.507	0.505	0.506	0.503	0.417	0.505
bias	0.001	0.004	0.007	0.005	0.006	0.003	-0.083	0.005
% bias	0.207	0.860	1.447	0.911	1.141	0.629	**-16.62**	0.935
model se	0.076	0.086	0.097	0.104	0.102	0.085	0.114	0.080
95% coverage	96.5	96.0	97.5	98.0	97.0	93.5	95.5	93.0
rmse	0.072	0.079	0.083	0.085	0.081	0.083	0.116	0.081
σ_{site}								
est	0.190	0.183	0.223	0.240	0.204	0.162	0.284	0.136
bias	-0.010	-0.017	0.023	0.040	0.004	-0.038	0.084	-0.064
% bias	-4.831	**-8.255**	**11.60**	**19.88**	1.940	**-19.02**	**42.15**	**-31.79**
σ_{error}								
est	0.981	0.982	0.981	0.994	0.992	0.984	1.012	0.992
bias	0.001	0.002	0.002	0.014	0.012	0.004	0.032	0.012
% bias	0.079	0.212	0.163	1.404	1.238	0.390	3.309	1.198
site ICC	0.038	0.037	0.051	0.057	0.043	0.029	0.074	0.020
average time to MI one data set (sec)			0.19	2.11	0.16	0.60	1.96	0.02

Table 3.3
Simulation results under the scenario with sporadically MCAR and systematically MAR outcome data. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC $= 0.04$.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
est	0.501	0.505	0.499	0.504	0.506	0.504	0.419	0.505
bias	0.001	0.005	-0.001	0.004	0.006	0.004	-0.081	0.005
% bias	0.207	0.912	-0.168	0.776	1.211	0.881	**-16.20**	0.982
model se	0.076	0.087	0.097	0.097	0.106	0.086	0.105	0.081
95% coverage	96.5	97.5	98.0	98.0	96.5	94.0	94.0	95.0
rmse	0.072	0.083	0.085	0.085	0.087	0.087	0.111	0.086
σ_{site}								
est	0.190	0.189	0.226	0.223	0.210	0.167	0.265	0.141
bias	-0.010	-0.011	0.026	0.023	0.010	-0.033	0.065	-0.059
% bias	-4.831	**-5.479**	**13.02**	**11.64**	4.779	**-16.46**	**32.69**	**-29.52**
σ_{error}								
est	0.981	0.980	0.978	0.981	0.993	0.981	0.980	0.989
bias	0.001	0.000	-0.002	0.002	0.013	0.001	0.000	0.010
% bias	0.079	0.005	-0.161	0.169	1.348	0.100	0.016	0.970
site ICC	0.038	0.039	0.052	0.051	0.045	0.031	0.070	0.022
average time to MI one data set (sec)			0.17	2.09	0.15	0.60	1.95	0.02

Table 3.4
Simulation results under the scenario with sporadically MAR and systematically MCAR outcome data. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
est	0.501	0.503	0.502	0.503	0.505	0.506	0.424	0.503
bias	0.001	0.003	0.002	0.003	0.005	0.006	-0.076	0.003
% bias	0.207	0.653	0.358	0.639	0.938	1.126	**-15.26**	0.649
model se	0.076	0.085	0.095	0.100	0.095	0.083	0.110	0.080
95% coverage	96.5	95.5	97.0	98.0	98.0	94.5	96.5	93.5
rmse	0.072	0.079	0.081	0.082	0.080	0.080	0.108	0.083
σ_{site}								
est	0.190	0.182	0.221	0.232	0.202	0.160	0.281	0.136
bias	-0.010	-0.018	0.021	0.032	0.002	-0.040	0.081	-0.064
% bias	-4.831	**-8.887**	**10.49**	**16.00**	0.813	**-20.11**	**40.64**	**-32.16**
σ_{error}								
est	0.981	0.982	0.981	0.992	0.991	0.983	1.013	0.991
bias	0.001	0.002	0.001	0.013	0.012	0.004	0.033	0.011
% bias	0.079	0.183	0.128	1.292	1.184	0.375	3.370	1.164
site ICC	0.038	0.036	0.050	0.053	0.042	0.028	0.073	0.020
average time to MI one data set (sec)			0.19	2.17	0.16	0.62	2.01	0.02

Table 3.5
Simulation results under the scenario with sporadically and systematically MCAR outcome data and a systematically MCAR covariate. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
est	0.501	0.500	0.495	0.495	0.495	0.400	0.494
bias	0.001	0.000	-0.005	-0.005	-0.005	-0.100	-0.006
% bias	0.207	0.044	-0.987	-1.091	-1.036	**-19.91**	-1.299
model se	0.076	0.118	0.104	0.099	0.087	0.107	0.085
95% coverage	96.5	95.5	97.5	96.5	95.5	93.5	94.0
rmse	0.072	0.117	0.081	0.083	0.083	0.123	0.083
σ_{site}							
est	0.190	0.179	0.229	0.203	0.155	0.263	0.133
bias	-0.010	-0.021	0.029	0.003	-0.045	0.063	-0.067
% bias	-4.831	**-10.47**	**14.45**	1.618	**-22.37**	**31.33**	**-33.60**
σ_{error}							
est	0.981	0.983	1.007	1.021	1.011	1.006	1.017
bias	0.001	0.003	0.027	0.041	0.031	0.026	0.037
% bias	0.079	0.344	2.767	4.196	3.145	2.635	3.780
site ICC	0.038	0.037	0.050	0.040	0.026	0.065	0.018
average time to MI one data set (sec)			0.33	0.28	0.83	2.96	0.03

Table 3.6
Simulation results under the scenario with sporadically and systematically MAR outcome data and a systematically MCAR covariate. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
est	0.501	0.507	0.497	0.496	0.498	0.399	0.499
bias	0.001	0.007	-0.003	-0.004	-0.002	-0.101	-0.001
% bias	0.207	1.351	-0.661	-0.706	-0.436	**-20.21**	-0.263
model se	0.076	0.118	0.105	0.108	0.087	0.117	0.084
95% coverage	96.5	95.5	98.0	97.5	94.5	94.0	95.0
rmse	0.072	0.123	0.087	0.086	0.082	0.128	0.084
σ_{site}							
est	0.190	0.175	0.227	0.208	0.151	0.287	0.126
bias	-0.010	-0.025	0.027	0.008	-0.049	0.087	-0.074
% bias	-4.831	**-12.60**	**13.37**	3.881	**-24.60**	**43.26**	**-36.84**
σ_{error}							
est	0.981	0.984	1.009	1.022	1.011	1.041	1.018
bias	0.001	0.004	0.030	0.042	0.032	0.062	0.038
% bias	0.079	0.401	3.031	4.306	3.227	**6.297**	3.909
site ICC	0.038	0.037	0.049	0.042	0.025	0.072	0.017
average time to MI one data set (sec)			0.37	0.30	0.88	3.20	0.04

Table 3.7
Simulation results under the scenario with sporadically MCAR and systematically MAR outcome data and a systematically MCAR covariate. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
est	0.501	0.500	0.497	0.499	0.497	0.398	0.497
bias	0.001	0.000	-0.003	-0.001	-0.003	-0.102	-0.003
% bias	0.207	0.061	-0.543	-0.239	-0.578	**-20.32**	-0.615
model se	0.076	0.119	0.105	0.111	0.089	0.109	0.086
95% coverage	96.5	94.0	97.5	98.0	96.5	90.0	95.5
rmse	0.072	0.124	0.087	0.092	0.087	0.126	0.088
σ_{site}							
est	0.190	0.179	0.230	0.206	0.156	0.268	0.132
bias	-0.010	-0.021	0.030	0.006	-0.044	0.068	-0.068
% bias	-4.831	**-10.64**	**15.13**	2.771	**-22.13**	**33.88**	**-33.88**
σ_{error}							
est	0.981	0.982	1.006	1.019	1.010	1.004	1.016
bias	0.001	0.002	0.026	0.040	0.030	0.024	0.036
% bias	0.079	0.213	2.671	4.039	3.109	2.492	3.713
site ICC	0.038	0.038	0.051	0.041	0.026	0.068	0.018
average time to MI one data set (sec)			0.33	0.29	0.86	3.07	0.03

Table 3.8
Simulation results under the scenario with sporadically MAR and systematically MCAR outcome data and a systematically MCAR covariate. The true values are $\beta_T = 0.50$, $\sigma_{site} = 0.2000$, $\sigma_{error} = 0.9798$, and ICC = 0.04.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
est	0.501	0.505	0.499	0.494	0.496	0.403	0.496
bias	0.001	0.005	-0.001	-0.006	-0.004	-0.097	-0.004
% bias	0.207	0.975	-0.125	-1.119	-0.882	**-19.47**	-0.720
model se	0.076	0.117	0.104	0.099	0.086	0.115	0.083
95% coverage	96.5	94.0	98.0	98.0	93.0	93.0	94.0
rmse	0.072	0.122	0.085	0.086	0.086	0.122	0.083
σ_{site}							
est	0.190	0.171	0.222	0.203	0.149	0.280	0.126
bias	-0.010	-0.029	0.022	0.003	-0.051	0.080	-0.074
% bias	-4.831	**-14.27**	**11.01**	1.382	**-25.70**	**40.16**	**-37.24**
σ_{error}							
est	0.981	0.983	1.008	1.020	1.012	1.040	1.017
bias	0.001	0.003	0.028	0.040	0.032	0.060	0.037
% bias	0.079	0.279	2.884	4.098	3.256	**6.094**	3.825
site ICC	0.038	0.036	0.048	0.040	0.023	0.069	0.017
average time to MI one data set (sec)			0.37	0.31	0.89	3.24	0.03

missing data. Smaller numbers of clusters tend to result in an underestimation of σ_{site} (Maas and Hox, 2004; Moerbeek and Teerenstra, 2015).

The FCS-LMM-MLE method outperformed the FCS-LMM and the FCS-LMM-het methods across all scenarios. While the FCS-GLM method returned coverage of the β_T estimate closest to 95%, the FCS-LMM-MLE method gave more accurate estimates of σ_{site}. The absolute percent bias for estimating σ_{site} was lowest and under 5% using FCS-LMM-MLE. This method also consistently took the shortest time to run among the multilevel MI methods. When introducing a systematically MCAR covariate, the FCS-LMM-het method did not run but the performance of the other methods did not appreciably change. Based on these findings and its other capabilities, we considered the FCS-LMM-MLE method as the best performing of the methods we evaluated for our purposes.

CHAPTER 4

Methods for Sensitivity Analysis

The MI methods described in Chapter 2 and evaluated in Chapter 3 assume that the missing data are MAR. When data are MNAR, these methods may give biased results. Therefore when there is uncertainty about the missing data mechanism, sensitivity analyses under MNAR assumptions should be conducted to evaluate the robustness of inferences (Little and Rubin, 2002; White et al., 2011a). In Section 4.1 we develop methods for conducting sensitivity analysis for CRTs with sporadically and systematically missing outcome data. The methods are then applied to the KHLP data in Section 4.2.

4.1 Methods for Handling MNAR Assumptions

The two common modeling frameworks for generating data under MNAR assumptions are selection models and pattern-mixture models (Harel and Zhou, 2007; van Buuren, 2018; Fiero et al., 2017). These two approaches decompose the joint distribution $P(\mathbf{Y}, \mathbf{R})$ in different ways. Selection models (Heckman, 1976) decompose the joint distribution as the marginal distribution of $\mathbf{Y}$ times the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}$, $P(\mathbf{Y}, \mathbf{R}) = P(\mathbf{Y})P(\mathbf{R} \,|\, \mathbf{Y})$. Both $P(\mathbf{Y})$ and $P(\mathbf{R} \,|\, \mathbf{Y})$ are unknown and must be specified. Selection models are sensitive to these specifications and require strong assumptions to describe potential dropout patterns. Pattern-mixture models (Glynn et al., 1986) decompose the joint distribution as the marginal distribution of $\mathbf{R}$ times the conditional distribution of $\mathbf{Y}$ given $\mathbf{R}$. The joint distribution can be expressed as a mixture of the distributions of $\mathbf{Y}$ for completers and dropouts:

$$\begin{aligned} P(\mathbf{Y}, \mathbf{R}) &= P(\mathbf{Y} \mid \mathbf{R}) P(\mathbf{R}) \\ &= P(\mathbf{Y} \mid \mathbf{R} = \mathbf{1}) P(\mathbf{R} = \mathbf{1}) + P(\mathbf{Y} \mid \mathbf{R} = \mathbf{0}) P(\mathbf{R} = \mathbf{0}). \end{aligned}$$

The probabilities $P(\mathbf{R} = \mathbf{1})$ and $P(\mathbf{R} = \mathbf{0})$ are the overall proportions of observed and missing data, respectively. The distribution $P(\mathbf{R} = \mathbf{1})$ of the completers can be modeled after the observed data, but the distribution $P(\mathbf{Y} \mid \mathbf{R} = \mathbf{0})$ of the dropouts needs to be specified.

Various selection and pattern-mixture modeling methods for handling MNAR data have been developed. While existing methods range in complexity, many authors have advocated for the use of simple and easily reproducible pattern-mixture model techniques to perform sensitivity analyses under MNAR assumptions (Rubin, 1987; Little, 2009; van Buuren, 2018). One such technique is introducing a shift parameter δ or a scale parameter k to imputed values generated under a MAR assumption to derive MNAR imputed values. If we suppose that the mean outcome among dropouts differs from that of completers by a constant δ, then δ can be added to the MAR imputed values. Similarly if we suppose that the mean outcome among dropouts and completers differs by some percentage $(k \times 100)\%$, we can multiply each MAR imputed value by $1 + k$. For example, if dropouts are suspected to have worse outcomes by 20% and higher outcomes are worse, then we would set $k = 0.2$ and multiply the imputed values by 1.2. This approach is attractive because it is a straightforward and accessible way of performing sensitivity analysis under a variety of MNAR assumptions. The impact of different assumptions on the study results is interpretable to both a statistical and non-statistical audience.

Within the framework of MI methods, researchers have applied δ- and k-adjustments to single-level data and longitudinal data (van Buuren, 2018; Leacy et al., 2017; Cro et al., 2020). Fiero et al. (2017) explored using multilevel MI and a k-adjustment to handle missing continuous outcomes in a longitudinal CRT. They used MICE to perform multilevel MI then multiplied imputed values by $1 + k$ to get MNAR imputed values:

$$(\text{MNAR imputed } Y_i) = (1 + k) \times (\text{MAR imputed } Y_i). \tag{4.1}$$

When MAR imputed values could be negative, a more general equation was used (Siddique et al., 2012):

$$(\text{MNAR imputed } Y_i) = [k \times |\,\text{MAR imputed } Y_i\,|\,] + \text{MAR imputed } Y_i. \tag{4.2}$$

For missing outcomes in the control arm, they specified $k = 0$, which assumes that the unobserved values of dropouts were similar to the observed values of trial completers. For missing outcomes in the intervention arm, varying values of k were used to represent a range of assumptions about dropouts. The researchers only explored sporadically missing data in their simulations and applications.

4.1.1 Handling MNAR Assumptions with Cluster-Level Dropout

Building on Fiero et al. (2017), we develop approaches for applying the k- and δ-adjustment methods to an outcome variable in a CRT with both sporadically and systematically missing values. Our approach allows for different MNAR assumptions for sporadically and systematically missing values and integrates these into a multistep process. We use the FCS-LMM-MLE multilevel MI method, which was the best performing method among the evaluated methods in our simulation study. For k-adjustments, the algorithm is:

1. Use the FCS-LMM-MLE method to impute missing values under a MAR assumption.

2. Apply a scale parameter at the cluster level: Multiply the imputed values of individuals whose missing outcomes are due to systematic missingness by a value of $(1 + k_{sys})$.

3. Apply a scale parameter at the individual level: Multiply the imputed values of individuals whose missing outcomes are due to sporadic missingness by a value of $(1+k_{spr})$.

4. Analyze each multiply-imputed data set using a LMM and combine the results using Rubin's rules.

To implement the δ-adjustment method, the same process can be used, but the second and third steps are replaced with:

2. Add a shift parameter at the cluster level: Add a value of δ_{sys} to the imputed values of individuals whose missing outcomes are due to systematic missingness.

3. Apply a shift parameter at the individual level: Add a value of δ_{spr} to the imputed values of individuals whose missing outcomes are due to sporadic missingness.

By distinguishing between steps 2 and 3, we are able to test how applying stronger or weaker MNAR assumptions to systematically missing data compared to the sporadically missing data affects inference.

Various assumptions about dropouts in each arm can be evaluated using this approach. For example, we could assume that all systematically missing outcomes values are MAR. For sporadically missing outcome values due to individual dropout, we could assume missing outcome values in the control arm are MAR and missing outcome values in the intervention are MNAR. Researchers applying this method can turn to subject matter experts to inform the range of values for k or δ. Additionally, one can use tipping point analysis to test a range of low and high values of k or δ until the statistical inference changes from what is observed under the MAR assumption (Yan et al., 2009; Liublinska and Rubin, 2014). We illustrate this using a real data application in Section 4.2.

4.2 Application to KHLP Study

Our main focus in this dissertation is on missing outcomes in cluster randomized trials. However, as noted in Section 1.2.2, our motivating example from the KHLP study also has systematically missing baseline values of the outcome variable. We therefore applied two analysis models to our motivating example: a model which includes baseline knowledge as a covariate (and therefore had substantial covariate missingness), and a reduced model that does not include the incomplete baseline knowledge variable as a covariate (and therefore had missing outcomes only). This allowed us to evaluate a model with both an imputed outcome and covariate and a model with only an imputed outcome. In both analysis models, knowledge score at follow-up was the outcome variable and the model included treatment assignment and a random intercept for church.

4.2.1 Application of FCS-LMM-MLE Method

The results for models with only an incomplete outcome and with an incomplete baseline covariate and outcome are in Table 4.1. For the model including baseline knowledge, FCS-LMM-MLE was used to impute both the incomplete covariate and the incomplete outcome. For the model not including baseline knowledge as a covariate, FCS-LMM-MLE was used to impute the incomplete outcome only. MAR was assumed. Two imputation models were considered. The first imputation model only included treatment assignment as a predictor (without auxiliary predictors). The second imputation model included treatment assignment and the following predictors: church size, church location, age, and marital status (with auxiliary predictors). Two participants with incomplete marital status data were excluded. These auxiliary predictors were included in the model because they were found to predict the follow-up knowledge score value or predict whether knowledge score was missing at follow-up (White et al., 2011b). A random intercept for church was also included in all imputation models. We generated 10 imputed data sets. We include results from complete-case analysis for comparison (excluding data from the contaminated churches which are considered missing).

The parameters of interest were the mean difference in knowledge score between the control and intervention arms at follow-up (β_T) and the variance components σ_{site} and σ_{error}. For the model with only incomplete follow-up knowledge data, complete-case analysis provided an estimate of β_T that was closer to the null of no intervention effect and had a smaller standard error compared to the FCS-LMM-MLE without auxiliary predictors results. When compared to the estimate under FCS-LMM-MLE with auxiliary predictors, complete-case analysis gave an estimate of β_T that was further from the null and had a smaller standard error. The variance component σ_{site} was smallest using FCS-LMM-MLE without auxiliary predictors as was the ICC. The estimate of σ_{site} and the ICC were largest using FCS-LMM-MLE with auxiliary predictors. An explanation for this may be that when site and demographic characteristics vary between sites and are included as predictors in the imputation model, site heterogeneity is increased. In the simulation study, the estimate of β_T and its standard error were larger using FCS-LMM-MLE. The variance component σ_{site} and the ICC were also larger using FCS-LMM-MLE.

For the model with incomplete baseline and follow-up knowledge data, complete-case analysis provided an estimate of β_T that was closer to the null of no intervention effect compared to the FCS-LMM-MLE results. The standard error of $\hat{\beta}_T$ was smallest when using FCS-LMM-MLE without auxiliary predictors and largest using FCS-LMM-MLE with auxiliary predictors. The variance component σ_{site} was smallest using complete-case analysis as was the ICC. Comparing the estimate of σ_{site} using FCS-LMM-MLE by imputation model, we see that including auxiliary predictors increases the estimate of σ_{site}. In the simulation study scenarios where both baseline knowledge and the outcome were missing, the estimate of β_T and its standard error were smaller using FCS-LMM-MLE compared to complete-case analysis. The same patterns for the variance components and the ICC were seen in the simulation study.

The FCS-LMM-MLE method was implemented using the *mice.impute.2l.continuous* function of the *miceadds* package in R. The run-times were about 10 and 20 seconds for the incomplete outcome model and incomplete outcome and covariate model, respectively.

Table 4.1
Results of the application of a LMM to the KHLP data using the complete-case (CC) and FCS-LMM-MLE methods.

	Incomplete follow-up knowledge data			**Incomplete baseline and follow-up knowledge data**		
	CC	**FCS-LMM-MLE**		**CC**	**FCS-LMM-MLE**	
		Without auxiliary predictors	With auxiliary predictors		Without auxiliary predictors	With auxiliary predictors
	$n_{individual} = 845$ $n_{cluster} = 49$	$n_{individual} = 1123$ $n_{cluster} = 52$	$n_{individual} = 1121$ $n_{cluster} = 52$	$n_{individual} = 845$ $n_{cluster} = 49$	$n_{individual} = 1123$ $n_{cluster} = 52$	$n_{individual} = 1121$ $n_{cluster} = 52$
β_T						
est	0.537	0.543	0.533	0.546	0.559	0.571
se	0.119	0.121	0.131	0.149	0.125	0.168
95% CI	(0.304, 0.769)	(0.305, 0.780)	(0.276, 0.790)	(0.253, 0.838)	(0.314, 0.804)	(0.242, 0.900)
σ_{site} est	0.203	0.181	0.234	0.184	0.207	0.350
σ_{error} est	1.429	1.428	1.435	1.318	1.325	1.366
site ICC	0.020	0.016	0.026	0.019	0.024	0.060

Abbreviations: est, estimate; se, standard error; CI, confidence interval

4.2.2 Sensitivity Analysis for KHLP Study

We conducted a sensitivity analysis with the KHLP study data to illustrate the process. The sensitivity analysis applied k-adjustments and δ-adjustments to the MAR imputed values from the FCS-LMM-MLE method with auxiliary predictors to create MNAR values. We allowed the value of the adjustment parameter to vary separately depending on whether the imputed data were sporadically missing (k_{spr} or δ_{spr}) or systematically missing (k_{sys} or δ_{sys}).

For this illustration, we assumed dropouts in the control arm were MAR and dropouts in the intervention arm were MNAR. In the KHLP study, the systematically missing values were due to contamination and are therefore presumably MAR in both arms. We considered a range of k_{spr} and k_{sys} or of δ_{spr} and δ_{sys} values that included a MAR value and increasingly strong MNAR assumptions. We were interested in finding the tipping point, or point at which inference for β_T changed from what was observed under MAR.

k-adjustment

Within the intervention arm, the MNAR assumption was that knowledge score at follow-up would be worse (lower) among dropouts compared to completers. Among individuals with sporadically missing outcome data, we decreased the imputed values in the intervention arm by increments of 10% from 0% to 50% corresponding to $k = (0, -0.1, -0.2, -0.3, -0.4, -0.5)$. The same range of values was considered for k_{spr} and k_{sys}.

The results of applying the k-adjustment are in Table 4.2 and Figure 4.1. For the model including only an imputed outcome, when systematically missing data were assumed to be MAR ($k_{sys} = 0$), the tipping point occurred when we assumed sporadic intervention dropouts had between 20% and 30% lower ($k_{spr} = -0.2$ and $k_{spr} = -0.3$) knowledge scores compared to completers. When we assumed that systematically missing data were also MNAR ($k_{sys} \neq 0$), the tipping point occurred at larger values of k_{spr}. Once we assumed the dropouts with sporadic missingness performed worse by 30% or more ($1 + k_{spr} \leq 0.7$) or that dropouts with systematically missing data performed worse by 30% or more ($1 + k_{sys} \leq 0.7$), a significant intervention effect was no longer observed. The same pattern was observed for the model with both imputed baseline and follow-up knowledge score data.

Table 4.2

Treatment effect β_T (95% CI) results of k-adjustment application to the KHLP data.

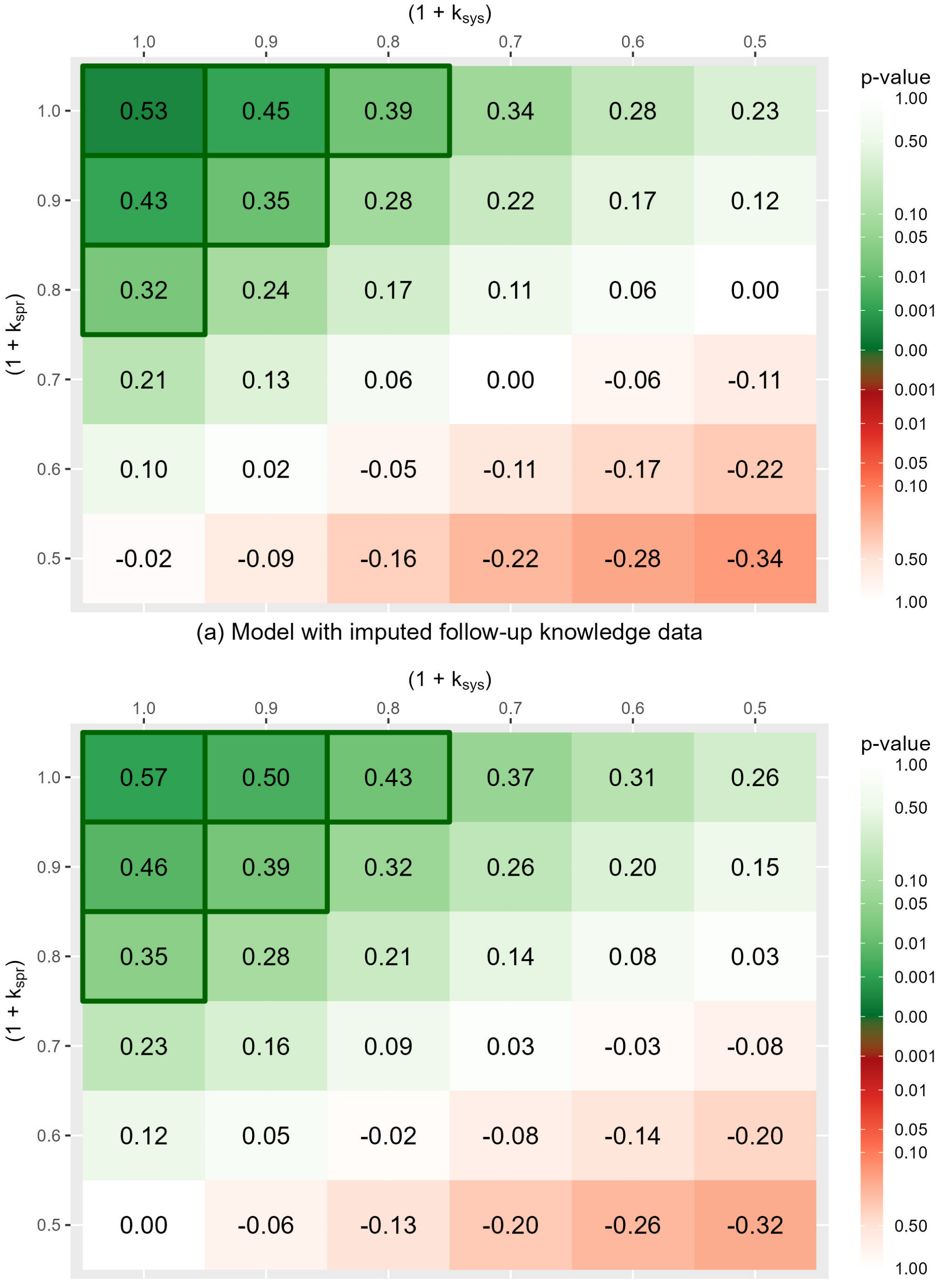
			Model with imputed follow-up data			
			$1 + k_{sys}$			
$1 + k_{spr}$	1.0	0.9	0.8	0.7	0.6	0.5
1.0	0.53 (0.28, 0.79)	0.45 (0.18, 0.73)	0.39 (0.07, 0.71)	0.34 (-0.03, 0.70)	0.28 (-0.14, 0.70)	0.23 (-0.24, 0.70)
0.9	0.43 (0.17, 0.69)	0.35 (0.07, 0.62)	0.28 (-0.03, 0.59)	0.22 (-0.14, 0.58)	0.17 (-0.24, 0.58)	0.12 (-0.35, 0.58)
0.8	0.32 (0.05, 0.59)	0.24 (-0.04, 0.52)	0.17 (-0.14, 0.48)	0.11 (-0.25, 0.47)	0.06 (-0.35, 0.46)	0.00 (-0.46, 0.46)
0.7	0.21 (-0.08, 0.50)	0.13 (-0.16, 0.42)	0.06 (-0.26, 0.38)	0.00 (-0.36, 0.36)	-0.06 (-0.47, 0.35)	-0.11 (-0.57, 0.35)
0.6	0.10 (-0.22, 0.41)	0.02 (-0.28, 0.33)	-0.05 (-0.38, 0.28)	-0.11 (-0.48, 0.26)	-0.17 (-0.59, 0.25)	-0.22 (-0.69, 0.24)
0.5	-0.02 (-0.36, 0.32)	-0.09 (-0.42, 0.25)	-0.16 (-0.51, 0.19)	-0.22 (-0.61, 0.16)	-0.28 (-0.71, 0.15)	-0.34 (-0.81, 0.14)
			Model with imputed baseline and follow-up data			
			$1 + k_{sys}$			
$1 + k_{spr}$	1.0	0.9	0.8	0.7	0.6	0.5
1.0	0.57 (0.24, 0.90)	0.50 (0.17, 0.82)	0.43 (0.08, 0.77)	0.37 (-0.01, 0.75)	0.31 (-0.11, 0.74)	0.26 (-0.22, 0.74)
0.9	0.46 (0.14, 0.78)	0.39 (0.07, 0.70)	0.32 (-0.02, 0.65)	0.26 (-0.12, 0.63)	0.20 (-0.22, 0.62)	0.15 (-0.32, 0.62)
0.8	0.35 (0.02, 0.67)	0.28 (-0.04, 0.59)	0.21 (-0.13, 0.54)	0.14 (-0.23, 0.51)	0.08 (-0.33, 0.50)	0.03 (-0.44, 0.50)
0.7	0.23 (-0.10, 0.57)	0.16 (-0.16, 0.49)	0.09 (-0.25, 0.44)	0.03 (-0.35, 0.41)	-0.03 (-0.45, 0.39)	-0.08 (-0.56, 0.39)
0.6	0.12 (-0.24, 0.47)	0.05 (-0.29, 0.40)	-0.02 (-0.37, 0.34)	-0.08 (-0.47, 0.31)	-0.14 (-0.57, 0.29)	-0.20 (-0.68, 0.28)
0.5	0.00 (-0.38, 0.38)	-0.06 (-0.43, 0.31)	-0.13 (-0.51, 0.25)	-0.20 (-0.60, 0.21)	-0.26 (-0.70, 0.19)	-0.32 (-0.81, 0.18)

δ-adjustment

Dropouts in the intervention arm were assumed to be MNAR with worse knowledge scores at follow-up compared to completers. We pegged values of δ to the observed standardized effect size among completers, which was 0.38 (calculated as $\hat{\beta}_T$ divided by the pooled standard deviation at follow-up of 1.44). Among individuals with sporadically missing outcome data, we decreased the imputed values in the intervention arm by increments of the standardized effect size (0.38) from 0 to 1.90 corresponding to $\delta = (0, -0.38, -0.76, -1.14, -1.52, -1.90)$. The same range of values was considered for δ_{spr} and δ_{sys}.

The results of applying the δ-adjustment are in Table 4.3 and Figure 4.2. For the model with only an imputed outcome, when systematically missing data were assumed to be MAR ($\delta_{sys} = 0$), the tipping point occurred when we assumed sporadic intervention dropouts had lower knowledge scores by between 1.52 and 1.90 standard deviation units ($\delta_{spr} = -1.52$ and $\delta_{spr} = -1.90$) compared to completers. When we assumed that systematically missing data were also MNAR ($\delta_{sys} \neq 0$), the tipping point occurred at larger values of δ_{spr}. Once we assumed the dropouts with sporadic missingness performed worse by 1.90 standard deviation units or more ($\delta_{spr} \leq -1.90$) or that dropouts with systematically missing data performed worse by 1.90 standard deviation units or more ($\delta_{sys} \leq -1.90$), a significant intervention effect was no longer observed. When the δ-adjustment method was applied to the model including an imputed baseline knowledge score covariate, a similar pattern was observed as when applied to the model not including baseline knowledge score. However, when systematically missing data were assumed to be MAR ($\delta_{sys} = 0$), the tipping point occurred when we assumed sporadic intervention dropouts had lower knowledge scores by between 1.14 and 1.52 standard deviation units ($\delta_{spr} = -1.14$ and $\delta_{spr} = -1.52$) compared to completers.

Figure 4.1:
Heat map representing p-values obtained by k-adjustment application, with corresponding treatment effects reported in each cell. Positive and negative treatment effects are denoted by the green and red gradients, respectively. The green grid highlights combinations that result in p-values < 0.05, with the staircase region indicating the tipping point.



(a) Model with imputed follow-up knowledge data

(b) Model with imputed baseline and follow-up knowledge data

Table 4.3
Treatment effect β_T (95% CI) results of δ-adjustment application to the KHLP data.

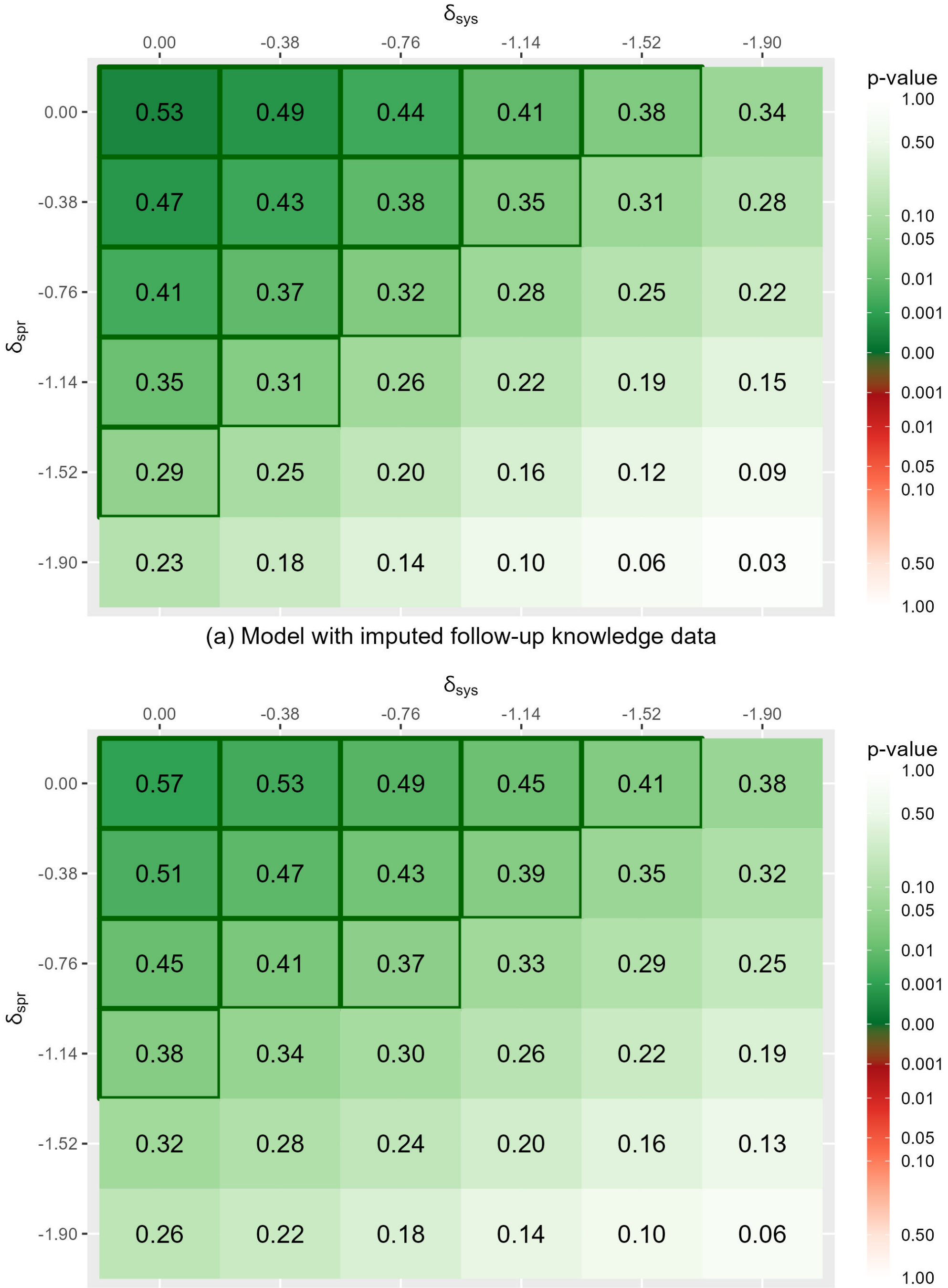
			Model with imputed follow-up data			
			δ_{sys}			
δ_{spr}	0.00	-0.38	-0.76	-1.14	-1.52	-1.90
0.00	0.53 (0.28, 0.79)	0.49 (0.22, 0.75)	0.44 (0.16, 0.73)	0.41 (0.10, 0.71)	0.38 (0.04, 0.71)	0.34 (-0.02, 0.70)
-0.38	0.47 (0.21, 0.73)	0.43 (0.16, 0.69)	0.38 (0.10, 0.67)	0.35 (0.04, 0.65)	0.31 (-0.02, 0.64)	0.28 (-0.08, 0.64)
-0.76	0.41 (0.15, 0.68)	0.37 (0.10, 0.64)	0.32 (0.04, 0.61)	0.28 (-0.02, 0.59)	0.25 (-0.08, 0.58)	0.22 (-0.14, 0.57)
-1.14	0.35 (0.08, 0.63)	0.31 (0.03, 0.58)	0.26 (-0.03, 0.55)	0.22 (-0.09, 0.53)	0.19 (-0.15, 0.52)	0.15 (-0.21, 0.51)
-1.52	0.29 (0.00, 0.57)	0.25 (-0.04, 0.53)	0.20 (-0.09, 0.50)	0.16 (-0.15, 0.47)	0.12 (-0.21, 0.46)	0.09 (-0.27, 0.45)
-1.90	0.23 (-0.07, 0.52)	0.18 (-0.11, 0.48)	0.14 (-0.16, 0.44)	0.10 (-0.22, 0.42)	0.06 (-0.28, 0.40)	0.03 (-0.34, 0.39)
			Model with imputed baseline and follow-up data			
			δ_{sys}			
δ_{spr}	0.00	-0.38	-0.76	-1.14	-1.52	-1.90
0.00	0.57 (0.24, 0.90)	0.53 (0.20, 0.86)	0.49 (0.15, 0.83)	0.45 (0.10, 0.80)	0.41 (0.04, 0.78)	0.38 (-0.01, 0.77)
-0.38	0.51 (0.18, 0.84)	0.47 (0.13, 0.80)	0.43 (0.09, 0.77)	0.39 (0.04, 0.74)	0.35 (-0.02, 0.72)	0.32 (-0.07, 0.71)
-0.76	0.45 (0.11, 0.78)	0.41 (0.07, 0.74)	0.37 (0.02, 0.71)	0.33 (-0.03, 0.68)	0.29 (-0.08, 0.66)	0.25 (-0.14, 0.64)
-1.14	0.38 (0.04, 0.73)	0.34 (-0.00, 0.69)	0.30 (-0.04, 0.65)	0.26 (-0.09, 0.62)	0.22 (-0.15, 0.60)	0.19 (-0.20, 0.58)
-1.52	0.32 (-0.03, 0.67)	0.28 (-0.07, 0.63)	0.24 (-0.11, 0.60)	0.20 (-0.16, 0.57)	0.16 (-0.21, 0.54)	0.13 (-0.27, 0.52)
-1.90	0.26 (-0.11, 0.62)	0.22 (-0.14, 0.58)	0.18 (-0.18, 0.54)	0.14 (-0.23, 0.51)	0.10 (-0.28, 0.48)	0.06 (-0.34, 0.46)

4.3 Discussion

Our k- and δ-adjustment sensitivity analysis methods are straightforward, accessible, and return easily interpretable results. The application to the KHLP study data showed how the two adjustment methods allow for the MNAR assumption to be conceptualized in different ways by using a scale or shift parameter. Using the scale parameter k, we saw that decreasing imputed values by increments of 10% returned a wider range of treatment effects across MNAR assumptions, with the most extreme assumptions returning negative treatment effects. The heat maps clearly illustrate that once dropouts (sporadic or systematic) in the intervention arm were assumed to perform worse than completers by 30% or more, the results are no longer significant. Using the shift parameter δ, even the most extreme MNAR assumptions returned positive treatment effects when the imputed values were decreased by increments of the observed standardized effect size of 0.38. The tipping point occurred when dropouts (sporadic or systematic) in the intervention arm were assumed to perform worse by 1.90 standard deviation units (5 times the observed standardized effect size) or more compared to completers. Given that it is extremely unlikely that the dropouts had such low scores, these findings suggest that under reasonable MNAR assumptions a positive and significant treatment effect would still be observed. Researchers can choose which adjustment method and what range of values to use based on the MNAR assumptions they want to evaluate and how they conceptualize the assumed difference between dropouts and completers.

Implementing the k- and δ-adjustment sensitivity analysis was simple and not computationally expensive. R code to execute the sensitivity analysis process is provided in Appendix B.

Figure 4.2:
Heat map representing p-values obtained by δ-adjustment application, with corresponding treatment effects reported in each cell. Positive and negative treatment effects are denoted by the green and red gradients, respectively. The green grid highlights combinations that result in p-values < 0.05, with the staircase region indicating the tipping point.



	0.00	-0.38	-0.76	-1.14	-1.52	-1.90
0.00	0.53	0.49	0.44	0.41	0.38	0.34
-0.38	0.47	0.43	0.38	0.35	0.31	0.28
-0.76	0.41	0.37	0.32	0.28	0.25	0.22
-1.14	0.35	0.31	0.26	0.22	0.19	0.15
-1.52	0.29	0.25	0.20	0.16	0.12	0.09
-1.90	0.23	0.18	0.14	0.10	0.06	0.03

(a) Model with imputed follow-up knowledge data

	0.00	-0.38	-0.76	-1.14	-1.52	-1.90
0.00	0.57	0.53	0.49	0.45	0.41	0.38
-0.38	0.51	0.47	0.43	0.39	0.35	0.32
-0.76	0.45	0.41	0.37	0.33	0.29	0.25
-1.14	0.38	0.34	0.30	0.26	0.22	0.19
-1.52	0.32	0.28	0.24	0.20	0.16	0.13
-1.90	0.26	0.22	0.18	0.14	0.10	0.06

(b) Model with imputed baseline and follow-up knowledge data

CHAPTER 5

Discussion

Missing outcome data are common in cluster randomized trials. While multilevel multiple imputation methods that account for clustering have become available, it has not been clear which methods might be best to apply when a CRT has data that are both sporadically and systematically missing (Audigier et al., 2018; Huque et al., 2020). Our simulation study results revealed that across the missing data mechanism scenarios and MI methods we considered, the FCS-LMM-MLE method performed best at handling an outcome with both sporadically and systematically missing data. While the FCS-LMM, FCS-LMM-het, FCS-GLM, and even the FCS-stnd methods performed similarly when estimating the intervention effect β_T, the FCS-LMM-MLE method was better at estimating σ_{site}, which is an especially important parameter for CRTs. It yielded the lowest absolute percent bias, which was under 5% across all scenarios.

The FCS-2stage method performed poorly in the scenarios that we considered. Audigier et al. (2018) showed that the FCS-2stage method returned biased results when there were small clusters in a data set. In particular, when imputing an incomplete continuous covariate, cluster sizes of 25 or less resulted in biased estimates of the regression coefficient. Estimates for the variance of random slope were biased when cluster sizes were less than 100. The simulation study we conducted had only 20 individuals per cluster. Hence the observed poor performance of the FCS-2stage method may have been due to small cluster sizes, and it is possible that the method would perform better under scenarios with larger cluster sizes.

Many authors have reported that the between-cluster variance tends to be underestimated when the number of clusters is small, with researchers defining "small" as anywhere from less than 50 to less than 100 (Busing, 1993; van der Leeden and Busing, 1994; van der Leeden et al., 1997; Maas and Hox, 2004). In the simulation study we conducted, 10% of the 60 clusters had systematically missing data. Complete case analysis using data from the 54 fully or partially observed clusters resulted in estimates of σ_{site} that were biased downward. The bias may be attributable to the smaller number of clusters involved in the complete case analysis.

We observed that for each method, the performance across the 4 missing data mechanism scenarios was similar. This suggests that the performance of any one method will not be substantially different depending on whether missing outcome data are sporadically and systematically MCAR, sporadically and systematically MAR, or MCAR for one type of missingness and MAR for the other. Audigier et al. (2018) previously found that the performance of FCS-GLM and FCS-2stage were similar whether sporadically missing covariate data were MCAR or MAR, but they did not consider scenarios where systematically missing data were MAR. Further, additionally imputing a missing covariate did not change the performance of the methods. An exception was the FCS-LMM-het method, which did not run in the presence of a systematically missing covariate.

Based on our findings, we recommend the use of the FCS-LMM-MLE method for a continuous outcome variable in a CRT with both sporadically and systematically missing values. Various authors have compared the performance of multilevel MI methods under different scenarios (Audigier et al., 2018; Grund et al., 2018; Huque et al., 2020), but to our knowledge the FCS-LMM-MLE method has not been evaluated in a similar way. FCS-LMM-MLE can be implemented using the *mice.impute.2l.continuous* function in the *miceadds* R package.

Application of the FCS-LMM-MLE method to the KHLP data showed that including auxiliary predictors in the imputation model increased the estimate of σ^2_{site}, the variance of the site-level random effect. The demographic characteristics of cluster members and site characteristics varied between sites in the KHLP data. It is well known that including covariates in a multilevel model often changes the variance components (Snijders and Bosker, 2011). Here, including covariates in the multilevel imputation model appears to have increased the estimate of σ_{site} in the multilevel analysis model. The association between auxiliary predictors and σ_{site} should be considered when selecting variables for imputation models and is an area that warrants further study.

We also developed approaches for conducting sensitivity analyses to evaluate the impact of missing data mechanism assumptions on inference. The multistep k-adjustment process that we describe facilitates sensitivity analysis under MNAR assumptions. A variety of MNAR assumptions can be tested by using a range of values for k. While k- and δ-adjustment methods have been applied to single level and longitudinal data (Liublinska and Rubin, 2014; Leacy et al., 2017; Cro et al., 2020), we applied the methods to CRT data. Our approach builds on Fiero et al. (2017) and allows the value of k to differ for systematically versus sporadically missing outcomes, allowing for different MNAR assumptions for these two types of missing data, which may have different mechanisms. We also describe a δ-adjustment process with the same flexibility. Although methods beyond k- and δ-adjustments have been developed to handle MNAR data (Carpenter et al., 2013; Galimard et al., 2016; Staudt et al., 2022), these adjustment methods are straightforward, do not require complex coding, and the impact of different MNAR assumptions on the results are easy to interpret. Liublinska and Rubin (2014) proposed graphical displays to visualize the results of sensitivity analyses. The heat maps we developed offer a visualization of the k- and δ-adjustment results. We restricted the MNAR assumptions to the intervention arm in our real data application. Other options include extending MNAR assumptions to the control arm and evaluating how allowing assumptions to vary by arms and by type of missingness affects results.

The imputation methods we considered can be applied generally to multilevel missing data, but more customized methods may be needed for some data structures. A two-stage

MI approach, distinct from the FCS-2stage method that we have discussed, was developed for situations where missing values fall into two qualitatively different types (Harel, 2007; Reiter and Raghunathan, 2007). The two-stage MI approach puts missing values into two groups and performs imputations by (1) imputing M_1 imputations of the first group's values and (2) imputing M_2 imputations of the second group's values given the imputed values from the first imputations. Harel (2007) suggested using two-stage MI to handle such situations as longitudinal studies with dropout and intermittent missingness and surveys where some questionnaires are fully unanswered whereas others are partially answered. Liu et al. (2016) adopted an MI approach based on additive regression, bootstrapping and predictive mean matching methods to handle sporadically and systematically missing accelerometer data. Standard MI methods were a poor option due to the repeated measures hierarchical data structure, complex patterns of missingness, and skewness of the data. More customized methods such as these approaches could be considered for missing CRT data when standard multilevel MI methods are not appropriate.

Further evaluation of how the MI methods presented here perform when applied to CRT data should be conducted. Only incomplete continuous outcomes and covariates were considered. The FCS-LMM-MLE, FCS-GLM, and FCS-2stage methods can be used for binary data. Future work could evaluate the performance of the methods in the presence of missing binary CRT data or a mix of missing continuous and binary CRT data. Other areas of future work could include comparing the performance of multilevel MI methods under different imputation and analytical models. The study design we simulated was balanced with 60 clusters and 20 individuals per cluster and true ICC of 0.04. Future work might evaluate the performance of the multilevel MI methods for unbalanced designs, smaller or larger sample sizes, or higher ICCs.

This dissertation serves to fill a gap in the literature on missing data in CRTs. While there is further research to be done, we identified FCS-LMM-MLE as an appropriate and well-performing MI method. We also developed a sensitivity analysis method that allows for different MNAR assumptions based on the whether data are sporadically or systematically missing.

APPENDIX A - MONTE CARLO STANDARD ERROR

MCSE results for the simulation study

Table A.1
MCSE results under the scenario with sporadically and systematically MCAR outcome data.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
bias	0.005	0.006	0.006	0.006	0.006	0.006	0.005	0.006
% bias	1.017	1.109	1.158	1.146	1.148	1.136	0.992	1.143
model se	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
95% coverage	1.300	1.386	0.990	1.206	1.386	1.612	1.541	1.541
rmse	0.023	0.026	0.026	0.026	0.026	0.027	0.028	0.027
σ_{site}								
bias	0.003	0.004	0.003	0.003	0.004	0.004	0.003	0.003
% bias	1.533	1.859	1.379	1.562	1.927	1.877	1.358	1.427
σ_{error}								
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.173	0.173	0.182	0.173	0.175	0.179	0.173

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.2
MCSE results under the scenario with sporadically and systematically MAR outcome data.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
bias	0.005	0.006	0.006	0.006	0.006	0.006	0.006	0.006
% bias	1.017	1.124	1.175	1.202	1.142	1.169	1.138	1.148
model se	0.000	0.001	0.001	0.001	0.002	0.001	0.001	0.001
95% coverage	1.300	1.386	1.104	0.990	1.206	1.743	1.466	1.804
rmse	0.023	0.025	0.026	0.026	0.025	0.026	0.032	0.025
σ_{site}								
bias	0.003	0.004	0.003	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.141	1.501	1.668	2.001	2.040	1.540	1.529
σ_{error}								
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.180	0.186	0.193	0.187	0.184	0.204	0.188

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.3
MCSE results under the scenario with sporadically MCAR and systematically MAR outcome data.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
bias	0.005	0.006	0.006	0.006	0.006	0.006	0.005	0.006
% bias	1.017	1.169	1.199	1.210	1.235	1.236	1.066	1.220
model se	0.000	0.001	0.001	0.001	0.002	0.001	0.001	0.001
95% coverage	1.300	1.104	0.990	0.990	1.300	1.679	1.679	1.541
rmse	0.023	0.025	0.025	0.026	0.026	0.026	0.031	0.026
σ_{site}								
bias	0.003	0.004	0.003	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.179	1.563	1.742	2.066	2.085	1.538	1.526
σ_{error}								
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.183	0.186	0.189	0.196	0.194	0.192	0.177

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.4
MCSE results under the scenario with sporadically MAR and systematically MCAR outcome data.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-het**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T								
bias	0.005	0.006	0.006	0.006	0.006	0.006	0.005	0.006
% bias	1.017	1.125	1.149	1.168	1.127	1.137	1.080	1.179
model se	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
95% coverage	1.300	1.466	1.206	0.990	0.990	1.612	1.300	1.743
rmse	0.023	0.023	0.024	0.025	0.024	0.024	0.031	0.024
σ_{site}								
bias	0.003	0.004	0.003	0.003	0.004	0.004	0.003	0.003
% bias	1.533	1.978	1.493	1.612	1.857	1.833	1.496	1.484
σ_{error}								
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.183	0.186	0.195	0.194	0.188	0.203	0.180

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.5
MCSE results under the scenario with sporadically and systematically MCAR outcome data and a systematically MCAR covariate.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
bias	0.005	0.008	0.006	0.006	0.006	0.005	0.006
% bias	1.017	1.656	1.153	1.174	1.179	1.019	1.172
model se	0.000	0.001	0.001	0.001	0.001	0.001	0.001
95% coverage	1.300	1.466	1.104	1.300	1.466	1.743	1.679
rmse	0.023	0.037	0.027	0.027	0.027	0.032	0.026
σ_{site}							
bias	0.003	0.005	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.659	1.357	1.939	1.922	1.447	1.501
σ_{error}							
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.240	0.183	0.194	0.190	0.193	0.188

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.6
MCSE results under the scenario with sporadically and systematically MAR outcome data and a systematically MCAR covariate.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
bias	0.005	0.009	0.006	0.006	0.006	0.006	0.006
% bias	1.017	1.735	1.233	1.216	1.163	1.106	1.184
model se	0.000	0.001	0.001	0.002	0.001	0.001	0.001
95% coverage	1.300	1.466	0.990	1.104	1.612	1.679	1.541
rmse	0.023	0.036	0.027	0.026	0.025	0.034	0.026
σ_{site}							
bias	0.003	0.006	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.868	1.500	2.044	2.064	1.585	1.591
σ_{error}							
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.250	0.187	0.193	0.188	0.209	0.192

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.7
MCSE results under the scenario with sporadically MCAR and systematically MAR outcome data and a systematically MCAR covariate.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
bias	0.005	0.009	0.006	0.007	0.006	0.005	0.006
% bias	1.017	1.758	1.234	1.304	1.230	1.063	1.247
model se	0.000	0.001	0.001	0.002	0.001	0.001	0.001
95% coverage	1.300	1.679	1.104	0.990	1.300	2.121	1.466
rmse	0.023	0.038	0.026	0.027	0.026	0.034	0.026
σ_{site}							
bias	0.003	0.006	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.773	1.527	2.032	1.999	1.554	1.604
σ_{error}							
bias	0.001	0.003	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.257	0.193	0.214	0.203	0.205	0.200

MCSEs of the % bias and 95% coverage parameters are reported as percentages

Table A.8
MCSE results under the scenario with sporadically MAR and systematically MCAR outcome data and a systematically MCAR covariate.

	Full	**CC**	**FCS-LMM**	**FCS-LMM-MLE**	**FCS-GLM**	**FCS-2stage**	**FCS-stnd**
β_T							
bias	0.005	0.009	0.006	0.006	0.006	0.005	0.006
% bias	1.017	1.734	1.200	1.215	1.224	1.032	1.170
model se	0.000	0.001	0.001	0.001	0.001	0.001	0.001
95% coverage	1.300	1.679	0.990	0.990	1.804	1.804	1.679
rmse	0.023	0.036	0.026	0.025	0.026	0.033	0.025
σ_{site}							
bias	0.003	0.006	0.003	0.004	0.004	0.003	0.003
% bias	1.533	2.838	1.432	1.773	1.847	1.457	1.470
σ_{error}							
bias	0.001	0.002	0.002	0.002	0.002	0.002	0.002
% bias	0.151	0.248	0.193	0.192	0.195	0.215	0.188

MCSEs of the % bias and 95% coverage parameters are reported as percentages

APPENDIX B - R CODE

R code to execute the sensitivity analysis process

```
## MI libraries
suppressMessages(library(dplyr))
suppressMessages(library(lme4))
suppressMessages(library(mice))
suppressMessages(library(miceadds))
suppressMessages(library(mitml))

## heatmap libraries
suppressMessages(library(ggplot2))
suppressMessages(library(tidyverse))
suppressMessages(library(latex2exp))
suppressMessages(library(RColorBrewer))

### Simulate KHLP Study data ###
set.seed(1130)

# site-level variables
site.vars <- data.frame(site = seq(1:60),
                        group = sample(rep(c("Intervention", "Control"),
                                           30)),
                        n_subs = rep(20, 60))

# subject-level data
# site ID
site <- 0
for (i in 1:60) {
  n_rows <- site.vars$n_subs[i]
  site_id.i <- rep(i, n_rows)
  site <- c(site, site_id.i)
}
site <- site[-1]

# subject ID
UniqueID <- "initiate"
for (i in 1:60) {
  n_rows <- site.vars$n_subs[i]
  kid_id.i <- c(1:n_rows)
  kid_id.i <- sprintf("%02d", kid_id.i)
  site.i <- sprintf("%03d", i)
  kid_id.i <- paste("C", site.i, kid_id.i, sep = "-")
  UniqueID <- c(UniqueID, kid_id.i)
}
UniqueID <- UniqueID[-1]

# sex, age, and baseline knowledge - based on KHLP Study distributions
sex <- sample(c("F", "M"), 1200, replace = TRUE, prob = c(.65, .35))
baseline_age <- rnorm(1200, 46.63844, 6.51258)
knowledge0 <- rnorm(1200, 6.05492, 1.49209)

dat.subj <- cbind.data.frame(site, UniqueID, sex,
                             baseline_age, knowledge0)
dat.subj <- merge(dat.subj, site.vars, by = "site", all.x = TRUE)

dat.subj <- dat.subj %>%
  arrange(site, UniqueID) %>%
  mutate(male = ifelse(sex == "M", 1, 0),
         intervention = ifelse(group == "Intervention", 1, 0)) %>%
  dplyr::select(UniqueID, site, baseline_age, knowledge0,
                sex, intervention)

# Generate BMI z-score data
# site effect
siteeff <- 999
for (i in 1:60) {
  n_rows <- site.vars$n_subs[i]
  site.ef.i <- rep(rnorm(1, mean = 0, sd = 0.04), n_rows)
  siteeff <- c(siteeff, site.ef.i)
}
siteeff <- siteeff[-1]
reseff = rnorm(1200, 0, 0.96)

knowledge6_complete <- 0.45 + 0.5*dat.subj[, "intervention"] +
  0.3*dat.subj[, "knowledge0"] + siteeff + reseff
dat.subj$knowledge6_complete <- knowledge6_complete

dat.subj.upd <- dat.subj %>%
  dplyr::select(UniqueID, site, knowledge0, knowledge6_complete,
                intervention, baseline_age, sex)
###############################################################


## Randomly delete follow-up measurements: MAR cluster/MCAR individual
# MAR: Drop follow-up for random ~10% of sites dependent on group
dat.site <- site.vars %>%
  mutate(control = ifelse(group == "Control", 1, 0))

x <- dat.site$control
logistic <- function(x) exp(x)/(1+exp(x))
p2.marright <- 1 - logistic(-2.75 + x)
r2.marright <- rbinom(dim(dat.site)[1], 1, p2.marright)
```

```
site.drop <- dat.site[r2.marright==0,]

sim.khlp <- dat.subj.upd %>%
  mutate(knowledge6 = ifelse(site %in% site.drop$site, NA,
                             knowledge6_complete))

dropped <- c(1)
n.drop <- length(site.drop$site)
for (i in 1:n.drop) {
  d <- site.drop$site[i]
  v <- c((d*20 - 19):(d*20))
  dropped <- c(dropped, v)
}
dropped <- dropped[-1]

# MCAR: Drop follow-up for random 20% of subjects
all.rows <- c(1:1200)
rm.rows <- all.rows[!all.rows %in% dropped]
subj.drop <- dat.subj.upd[sample(rm.rows, 240), ]

sim.khlp <- sim.khlp %>%
  mutate(knowledge6 = ifelse(UniqueID %in% subj.drop$UniqueID, NA,
                             knowledge6))
##############################################################


### Multiple Imputation ###
# Create data set for imputation
khlp.imp <- sim.khlp %>%
  mutate(missing = ifelse(is.na(knowledge6), 1, 0)) %>%
  group_by(site) %>%
  mutate(clst_missing = ifelse(sum(is.na(knowledge6)) == 20, 1, 0)) %>%
  dplyr::select(UniqueID, site, intervention, knowledge6,
                missing, clst_missing)

# FCS-GLM: group as predictor
predMatrix <- make.predictorMatrix(data = khlp.imp)
impMethod <- make.method(data = khlp.imp)

# method for outcome variable
impMethod[c("knowledge6")] <- "2l.continuous"

# remove indicator variables from predictor matrix
predMatrix[, c("UniqueID", "missing", "clst_missing")] <- 0
# specify cluster indicator
predMatrix[, "site"] <- -2

# specify cluster indicators
cluster <- list()
cluster[["knowledge6"]] <- c("site")

# Imputation
# run mice
imp.grp <- mice(khlp.imp, method = impMethod,
                 predictorMatrix = predMatrix, maxit = 20, m = 10,
                 levels_id = cluster, printFlag = FALSE)

# Analysis
# Fit model
implist.grp <- mids2mitml.list(imp.grp)
imp.fit <- lapply(implist.grp, FUN=function(x){
  lmer(knowledge6 ~ intervention + (1|site),
       data = x, REML = TRUE, control = lmerControl(optimizer ="bobyqa"))
})
```

```
# pool results
testEstimates(imp.fit, var.comp = TRUE)

##############################################################



### Sensitivity Analysis: k-adjustment and delta-adjustment ###
## k-adjustment: vary by level of missingness
range.k <- rev(seq(-.5, 0, by = .1))
m <- 10
n.k <- length(range.k)

# list of imputed data sets as one data frame
for (i in 1:m){
  implist.grp[[i]]$imp_set = i
  imp.i <- implist.grp[[i]]

  if(i == 1){
    imp.all <- imp.i
  } else {
    imp.all <- rbind(imp.all, imp.i)
  }
}

# data set of imputed data sets
imp.mnar <- imp.all %>%
  rename(knowledge = knowledge6) %>%
  mutate(indv_missing = ifelse(missing == 1 & clst_missing == 0, 1, 0))

# function to set various values of k
k_value_assign <- function(data, k_sys, k_spr){
  imp.dat <- data
  k_val <- ifelse(imp.dat$intervention == 1 &
                    imp.dat$clst_missing == 1, k_sys,
                  ifelse(imp.dat$intervention == 1 &
                           imp.dat$indv_missing == 1, k_spr, 0))
}

# loop through all values of k_sys and k_spr to create all offset values
for (i in 1:n.k){
  for (j in 1:n.k) {
    k_adj <- paste0("k", j-1, ".", i-1)
    imp.mnar[[k_adj]] <- k_value_assign(imp.mnar, range.k[i], range.k[j])
  }
}

# loop through offset values to create all MNAR imputed values
for (i in 1:n.k){
  for (j in 1:n.k) {
    k_adj <- paste0("k", j-1, ".", i-1)
    knw.mnar <- paste0("knowledge", j-1, ".", i-1)
    imp.mnar[[knw.mnar]] <- abs(imp.mnar$knowledge)*imp.mnar[[k_adj]] +
      imp.mnar$knowledge
  }
}

imp.mnar <- imp.mnar %>%
  dplyr::select(-c(k0.0:k5.5))

implist.grp.mnar <- list()
for (i in 1:m){
  implist.grp.mnar[[i]] <- imp.mnar %>% filter(imp_set == i)
}
```

```r
# Function: fit model
FUN.mnar <- function(dep.vars){
  N <- length(dep.vars)
  for(i in 1:N){
    imp.fit.mnar <- lapply(implist.grp.mnar, FUN=function(x){
      lmer(x[[dep.vars[i]]] ~ intervention + (1|site),
           data = x, REML = TRUE,
           control = lmerControl(optimizer ="bobyqa"))
    })
    est <- as.data.frame(testEstimates(imp.fit.mnar)$est[c(2), c(1,2,5)])
    est.1 <- sprintf("%.2f", round(est[1,1], 3))
    ci <- sprintf("%.2f", round(c(est[1,1]-(1.96*est[2,1]),
                                    est[1,1]+(1.96*est[2,1])), 3))
    diff.1 <- paste(est.1, " (", ci[1], ", ", ci[2], ")", sep = "")
    diff <- data.frame(grp_diff = est.1,
                       ci = paste("(", ci[1], ", ", ci[2], ")", sep = ""),
                       diff_ci = diff.1,
                       p_val = est[3,1])
    if(i == 1){
      diff.tab <- diff
    }
    else{
      diff.tab <- rbind(diff.tab, diff)
    }
    i <- i + 1
  }
  diff.tab
}

# vector of all variable names
var.names <- colnames(imp.mnar)
```

```
# cluster-level MAR: k = 0
knowledge.k0 <- var.names[grepl("knowledge\\d{1}.0", var.names)]
tab.grp.mnar.0 <- FUN.mnar(knowledge.k0)

# cluster-level MNAR: k = k1
knowledge.k1 <- var.names[grepl("knowledge\\d{1}.1", var.names)]
tab.grp.mnar.1 <- FUN.mnar(knowledge.k1)

# cluster-level MNAR: k = k2
knowledge.k2 <- var.names[grepl("knowledge\\d{1}.2", var.names)]
tab.grp.mnar.2 <- FUN.mnar(knowledge.k2)

# cluster-level MNAR: k = k3
knowledge.k3 <- var.names[grepl("knowledge\\d{1}.3", var.names)]
tab.grp.mnar.3 <- FUN.mnar(knowledge.k3)

# cluster-level MNAR: k = k4
knowledge.k4 <- var.names[grepl("knowledge\\d{1}.4", var.names)]
tab.grp.mnar.4 <- FUN.mnar(knowledge.k4)

# cluster-level MNAR: k = k5
knowledge.k5 <- var.names[grepl("knowledge\\d{1}.5", var.names)]
tab.grp.mnar.5 <- FUN.mnar(knowledge.k5)

# table of group differences and p-values under each MNAR assumption
tabk.grp.mnar <- cbind(tab.grp.mnar.0[,1], tab.grp.mnar.1[,1],
                          tab.grp.mnar.2[,1], tab.grp.mnar.3[,1],
                          tab.grp.mnar.4[,1], tab.grp.mnar.5[,1])
tabk.grp.mnar.p <- cbind(tab.grp.mnar.0[,4], tab.grp.mnar.1[,4],
                          tab.grp.mnar.2[,4], tab.grp.mnar.3[,4],
                          tab.grp.mnar.4[,4], tab.grp.mnar.5[,4])
```

```
## delta-adjustment: vary by level of missingness
range.d <- rev(seq(-1.90, 0, by = .38))
m <- 10
n.d <- length(range.d)

# list of imputed data sets as one data frame
for (i in 1:m){
  implist.grp[[i]]$imp_set = i
  imp.i <- implist.grp[[i]]

  if(i == 1){
    imp.all <- imp.i
  } else {
    imp.all <- rbind(imp.all, imp.i)
  }
}

# data set of imputed data sets
imp.mnar <- imp.all %>%
  rename(knowledge = knowledge6) %>%
  mutate(indv_missing = ifelse(missing == 1 & clst_missing == 0, 1, 0))

# function to set various values of delta
d_value_assign <- function(data, d_sys, d_spr){
  imp.dat <- data
  d_val <- ifelse(imp.dat$intervention == 1 &
                    imp.dat$clst_missing == 1, d_sys,
                  ifelse(imp.dat$intervention == 1 &
                           imp.dat$indv_missing == 1, d_spr, 0))
}
```

```
# loop through all values of d_sys and d_spr to create all offset values
for (i in 1:n.d){
  for (j in 1:n.d) {
    d_adj <- paste0("d", j-1, ".", i-1)
    imp.mnar[[d_adj]] <- d_value_assign(imp.mnar, range.d[i], range.d[j])
  }
}

# loop through offset values to create all MNAR imputed values
for (i in 1:n.d){
  for (j in 1:n.d) {
    d_adj <- paste0("d", j-1, ".", i-1)
    knw.mnar <- paste0("knowledge", j-1, ".", i-1)
    imp.mnar[[knw.mnar]] <- imp.mnar$knowledge + imp.mnar[[d_adj]]
  }
}

imp.mnar <- imp.mnar %>%
  dplyr::select(-c(d0.0:d5.5))

implist.grp.mnar <- list()
for (i in 1:m){
  implist.grp.mnar[[i]] <- imp.mnar %>% filter(imp_set == i)
}

# vector of all variable names
var.names <- colnames(imp.mnar)

# cluster-level MAR: d = 0
knowledge.d0 <- var.names[grepl("knowledge\\d{1}.0", var.names)]
tab.grp.mnar.0 <- FUN.mnar(knowledge.d0)
```

```
# cluster-level MNAR: d = d1
knowledge.d1 <- var.names[grepl("knowledge\\d{1}.1", var.names)]
tab.grp.mnar.1 <- FUN.mnar(knowledge.d1)

# cluster-level MNAR: d = d2
knowledge.d2 <- var.names[grepl("knowledge\\d{1}.2", var.names)]
tab.grp.mnar.2 <- FUN.mnar(knowledge.d2)

# cluster-level MNAR: d = d3
knowledge.d3 <- var.names[grepl("knowledge\\d{1}.3", var.names)]
tab.grp.mnar.3 <- FUN.mnar(knowledge.d3)

# cluster-level MNAR: d = d4
knowledge.d4 <- var.names[grepl("knowledge\\d{1}.4", var.names)]
tab.grp.mnar.4 <- FUN.mnar(knowledge.d4)

# cluster-level MNAR: d = d5
knowledge.d5 <- var.names[grepl("knowledge\\d{1}.5", var.names)]
tab.grp.mnar.5 <- FUN.mnar(knowledge.d5)

# table of group differences and p-values under each MNAR assumption
tabd.grp.mnar <- cbind(tab.grp.mnar.0[,1], tab.grp.mnar.1[,1],
                          tab.grp.mnar.2[,1], tab.grp.mnar.3[,1],
                          tab.grp.mnar.4[,1], tab.grp.mnar.5[,1])
tabd.grp.mnar.p <- cbind(tab.grp.mnar.0[,4], tab.grp.mnar.1[,4],
                          tab.grp.mnar.2[,4], tab.grp.mnar.3[,4],
                          tab.grp.mnar.4[,4], tab.grp.mnar.5[,4])


## Heat map: k-adjustment
# Function: convert to tibble, add row identifier, and shape "long"
```

```
FUN.df.plot_kadj <- function(data, parameter){
  df.plot <-
    data %>%
    as_tibble() %>%
    rownames_to_column("k_spr") %>%
    pivot_longer(-k_spr, names_to = "k_sys", values_to = parameter) %>%
    mutate(k_spr = factor(k_spr, levels = 1:10),
           k_sys = factor(gsub("V", "", k_sys), levels = 1:10)) %>%
    mutate(k_spr_num = 1 - (as.double(k_spr) - 1)/10,
           k_sys_num = 1 - (as.double(k_sys) - 1)/10) %>%
    mutate(k_spr = factor(sprintf("%.1f", round(k_spr_num, 1))),
           k_sys = fct_rev(factor(sprintf("%.1f", round(k_sys_num, 1)))))
  df.plot
}

df.tx <- FUN.df.plot_kadj(tabk.grp.mnar, "tx_ef")
df.p <- FUN.df.plot_kadj(tabk.grp.mnar.p, "p_val")

df.plot <- merge(df.tx, df.p,
                 by = c("k_spr", "k_sys", "k_spr_num", "k_sys_num"))
df.plot <- df.plot %>%
  arrange(desc(k_spr_num), desc(k_sys_num)) %>%
  group_by(k_sys) %>%
  mutate(tip_pnt = ifelse(p_val > 0.05 & lag(p_val) < 0.05, 1, 0)) %>%
  mutate(tip_pnt = ifelse(is.na(tip_pnt), 0, tip_pnt))

# Function: heat map for k-adjustment results
FUN.plot.border_kadj <- function(data.plot, pos.color, neg.color,
                                 border.pos, border.neg,
                                 level_sys, level_spr){
  heatmap.plot <-
    ggplot() +
    # fill for positive difference
    # when p-value >= .05
    geom_tile(data = {data.plot %>% filter(tx_ef >= 0 & p_val >= 0.05)},
              aes(x = k_sys, y = k_spr,
                  # scale p-values for better legend readability
                  fill = p_val^0.2676)) +
    # when p-value < .05 add tile border
    geom_tile(data = {data.plot %>% filter(tx_ef >= 0 & p_val < 0.05)},
              aes(x = k_sys, y = k_spr,
                  # scale p-values for better legend readability
                  fill = p_val^0.2676),
              color = border.pos, size = 1.2) +
    geom_text(data = {data.plot %>% filter(tx_ef >= 0)},
              aes(x = k_sys, y = k_spr,
                  label = sprintf("%.2f", as.double(tx_ef) + 0)),
              size = 4.5) +

    # fill for negative difference
    # when p-value >= .05
    geom_tile(data = {data.plot %>% filter(tx_ef < 0 & p_val >= 0.05)},
              aes(x = k_sys, y = k_spr,
                  # scale p-values for better legend readability
                  fill = -(p_val^0.2676))) +
    # when p-value < .05 add tile border
    geom_tile(data = {data.plot %>% filter(tx_ef < 0 & p_val < 0.05)},
              aes(x = k_sys, y = k_spr,
                  # scale p-values for better legend readability
                  fill = p_val^0.2676),
              color = border.neg, size = 1.2) +
    geom_text(data = {data.plot %>% filter(tx_ef < 0)},
              aes(x = k_sys, y = k_spr,
                  label = sprintf("%.2f", as.double(tx_ef) + 0)),
              size = 4.5) +
    # scale from 1 to 1
    scale_fill_gradientn(name = "p-value",
                         breaks = c(-c(1, .5, .1, 0.05,
                                       0.01, 0.001)^0.2676,
                                    c(0, 0.001, 0.01,
                                      0.05, 0.1, 0.5, 1)^0.2676),
                         labels = c("1.00", "0.50", "0.10", "0.05",
                                    "0.01", "0.001", "0.00",
                                    "0.001", "0.01", "0.05",
                                    "0.10", "0.50", "1.00"),
                         limits = c(-1, 1),
                         colors = c(neg.color, pos.color),
                         guide = guide_colorbar(barwidth = 0.8,
                                                barheight = 18)) +
    scale_x_discrete(position = "top",
                     limits = level_sys) +
    scale_y_discrete(limits = level_spr) +
    ylab(TeX("(1 + $k_{spr}$)")) +
    xlab(TeX("(1 + $k_{sys}$)"))

  heatmap.plot
}

# set colors
pos.color <- c(rev(brewer.pal(6, "Greens")), "#FFFFFF")
neg.color <- c("#FFFFFF", brewer.pal(5, "Reds"))

# set levels for k values (order in which they appear in graph)
level_spr <- levels(df.plot$k_spr)
level_sys <- levels(df.plot$k_sys)
```

```
FUN.plot.border_kadj(df.plot, pos.color, neg.color,
                     "darkgreen", "darkred", level_sys, level_spr)


## Heat map: delta-adjustment
# Function: convert to tibble, add row identifier, and shape "long"
FUN.df.plot_dadj <- function(data, parameter, d_int){
  df.plot <-
    data %>%
    as_tibble() %>%
    rownames_to_column("d_spr") %>%
    pivot_longer(-d_spr, names_to = "d_sys", values_to = parameter) %>%
    mutate(d_spr = factor(d_spr, levels = 1:10),
           d_sys = factor(gsub("V", "", d_sys), levels = 1:10)) %>%
    mutate(d_spr_num = (as.double(d_spr) - 1)*-d_int,
           d_sys_num = (as.double(d_sys) - 1)*-d_int) %>%
    mutate(d_spr = fct_rev(factor(sprintf("%.2f",
                                          round(d_spr_num, 2) + 0))),
           d_sys = factor(sprintf("%.2f", round(d_sys_num, 2) + 0)))
  df.plot
}

df.tx <- FUN.df.plot_dadj(tabd.grp.mnar, "tx_ef", 0.38)
df.p <- FUN.df.plot_dadj(tabd.grp.mnar.p, "p_val", 0.38)

df.plot <- merge(df.tx, df.p,
                 by = c("d_spr", "d_sys", "d_spr_num", "d_sys_num"))

df.plot$d_sys <- factor(df.plot$d_sys,
                        levels = c("0.00", levels(df.plot$d_sys)[-10]))
df.plot$d_spr <- factor(df.plot$d_spr,
                        levels = c(levels(df.plot$d_spr)[-1], "0.00"))
```

```
df.plot <- df.plot %>%
  arrange(desc(d_spr_num), desc(d_sys_num)) %>%
  group_by(d_sys) %>%
  mutate(tip_pnt = ifelse(p_val > 0.05 & lag(p_val) < 0.05, 1, 0)) %>%
  mutate(tip_pnt = ifelse(is.na(tip_pnt), 0, tip_pnt))

# Function: heat map for delta-adjustment results
FUN.plot.border_dadj <- function(data.plot, pos.color, neg.color,
                                 border.pos, border.neg,
                                 level_sys, level_spr){
  heatmap.plot <-
    ggplot() +
    # fill for positive difference
    # when p-value < .05 add tile border
    geom_tile(data = {data.plot %>% filter(tx_ef >= 0 & p_val < 0.05)},
              aes(x = d_sys, y = d_spr,
                  # scale p-values for better legend readability
                  fill = p_val^0.2676),
              color = border.pos, size = 1.2) +
    # when p-value >= .05
    geom_tile(data = {data.plot %>% filter(tx_ef >= 0 & p_val >= 0.05)},
              aes(x = d_sys, y = d_spr,
                  # scale p-values for better legend readability
                  fill = p_val^0.2676)) +
    geom_text(data = {data.plot %>% filter(tx_ef >= 0)},
              aes(x = d_sys, y = d_spr,
                  label = sprintf("%.2f", as.double(tx_ef) + 0)),
              size = 4.5) +

    # fill for negative difference
    # when p-value >= .05
```

```
geom_tile(data = {data.plot %>% filter(tx_ef < 0 & p_val >= 0.05)},
          aes(x = d_sys,
              y = d_spr,
              # scale p-values for better legend readability
              fill = -(p_val^0.2676))) +
# when p-value < .05 add tile border
geom_tile(data = {data.plot %>% filter(tx_ef < 0 & p_val < 0.05)},
          aes(x = d_sys,
              y = d_spr,
              # scale p-values for better legend readability
              fill = p_val^0.2676),
          color = border.neg, size = 1.2) +
geom_text(data = {data.plot %>% filter(tx_ef < 0)},
          aes(x = d_sys,
              y = d_spr,
              label = sprintf("%.2f", as.double(tx_ef) + 0)),
          size = 4.5) +
# scale from 1 to 1
scale_fill_gradientn(name = "p-value",
                     breaks = c(-c(1, .5, .1,
                                   0.05, 0.01, 0.001)^0.2676,
                                c(0, 0.001, 0.01,
                                  0.05, 0.1, 0.5, 1)^0.2676),
                     labels = c("1.00", "0.50", "0.10", "0.05",
                                "0.01", "0.001", "0.00",
                                "0.001", "0.01", "0.05",
                                "0.10", "0.50", "1.00"),
                     limits = c(-1, 1),
                     colors = c(neg.color, pos.color),
                     guide = guide_colorbar(barwidth = 0.8,
                                            barheight = 18)) +
scale_x_discrete(position = "top",
                     limits = level_sys) +
    scale_y_discrete(limits = level_spr) +
    ylab(TeX(r'(${\delta}_{spr}$)')) +
    xlab(TeX(r'(${\delta}_{sys}$)'))

  heatmap.plot
}

# set colors
pos.color <- c(rev(brewer.pal(6, "Greens")), "#FFFFFF")
neg.color <- c("#FFFFFF", brewer.pal(5, "Reds"))

# set levels for delta values (order in which they appear in graph)
level_spr <- levels(df.plot$d_spr)
level_sys <- levels(df.plot$d_sys)[c(6, 1:5)]

FUN.plot.border_dadj(df.plot, pos.color, neg.color,
                     "darkgreen", "darkred", level_sys, level_spr)
```

BIBLIOGRAPHY

Acosta, J., Chinman, M., Ebener, P., Malone, P. S., Phillips, A., and Wilks, A. (2019). Evaluation of a whole-school change intervention: Findings from a two-year cluster-randomized trial of the restorative practices intervention. *Journal of Youth and Adolescence*, 48:876–890.

Audigier, V. and Resche-Rigon, M. (2021). *micemd: Multiple Imputation by Chained Equations with Multilevel Data*. R package version 1.8.0.

Audigier, V., White, I. R., Jolani, S., Debray, T. P., Quartagno, M., Carpenter, J., van Buuren, S., and Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, 33(2):160–183.

Bastani, R. (2018). *Addressing Obesity in Early Care and Education Settings*. National Institutes of Health.

Bastani, R., Glenn, B. A., Maxwell, A. E., Jo, A. M., Herrmann, A. K., Crespi, C. M., Wong, W. K., Chang, L. C., Stewart, S. L., Nguyen, T. T., et al. (2015). Cluster-randomized trial to increase hepatitis B testing among Koreans in Los Angeles. *Cancer Epidemiology, Biomarkers & Prevention*, 24(9):1341–1349.

Braun, J. M., Kalkwarf, H. J., Papandonatos, G. D., Chen, A., and Lanphear, B. P. (2018). Patterns of early life body mass index and childhood overweight and obesity status at eight years of age. *BMC Pediatrics*, 18(1):1–8.

Busing, F. (1993). *Distribution characteristics of variance estimates in two-level models*. Unpublished manuscript. Department of Psychometrics and Research Methodology, Leiden University.

Carpenter, J. R., Roger, J. H., and Kenward, M. G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6):1352–1371.

Carpenter, J. R. and Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5):915–947.

Cro, S., Morris, T. P., Kenward, M. G., and Carpenter, J. R. (2020). Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*, 39(21):2815–2842.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.

Enders, C. K. (2022). *Applied Missing Data Analysis*. Guilford Publications.

Fiero, M. H., Hsu, C.-H., and Bell, M. L. (2017). A pattern-mixture model approach for handling missing continuous outcome data in longitudinal cluster randomized trials. *Statistics in Medicine*, 36(26):4094–4105.

Fiero, M. H., Huang, S., Oren, E., and Bell, M. L. (2016). Statistical analysis and handling of missing data in cluster randomized trials: A systematic review. *Trials*, 17:1–10.

Galimard, J.-E., Chevret, S., Protopopescu, C., and Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, 35(17):2907–2920.

Gasparini, A. (2018). rsimsum: Summarise results from monte carlo simulation studies. *Journal of Open Source Software*, 3:739.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In Wainer, H., editor, *Drawing Inferences from Self-Selected Samples*, pages 115–142. Springer.

Grund, S., Lüdtke, O., and Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, 21(1):111–149.

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4(1):75–89.

Harel, O. and Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16):3057–3077.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In Berg, S. V., editor, *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER.

Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B., and Lee, K. J. (2020). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal*, 62(2):444–466.

Isensee, B., Morgenstern, M., Stoolmiller, M., Maruska, K., Sargent, J. D., and Hanewinkel, R. (2012). Effects of Smokefree Class Competition 1 year after the end of intervention: A cluster randomised controlled trial. *Journal of Epidemiology & Community Health*, 66(4):334–341.

Jolani, S. (2018). Hierarchical imputation of systematically and sporadically missing data: An approximate Bayesian approach using chained equations. *Biometrical Journal*, 60(2):333–351.

Kasim, R. M. and Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, 23(2):93–116.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Leacy, F. P., Floyd, S., Yates, T. A., and White, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American Journal of Epidemiology*, 185(4):304–315.

Li, P., Stuart, E. A., and Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *JAMA*, 314(18):1966–1967.

Little, R. J. (2009). Comments on: Missing data methods in longitudinal studies: A review. *Test*, 18(1):47–50.

Little, R. J., Carpenter, J. R., and Lee, K. J. (2022). A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research*. 00491241221113873.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, B., Yu, M., Graubard, B. I., Troiano, R. P., and Schenker, N. (2016). Multiple imputation of completely missing repeated measures data within person from a complex sample: Application to accelerometer data in the National Health and Nutrition Examination Survey. *Statistics in Medicine*, 35(28):5170–5188.

Liublinska, V. and Rubin, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in Medicine*, 33(24):4170–4185.

Maas, C. J. and Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2):127–137.

Mendoza, J. A., McLeod, J., Chen, T.-A., Nicklas, T. A., and Baranowski, T. (2014). Correlates of adiposity among Latino preschool children. *Journal of Physical Activity & Health*, 11(1).

Moberg, J. and Kramer, M. (2015). A brief history of the cluster randomised trial design. *Journal of the Royal Society of Medicine*, 108(5):192–198.

Moerbeek, M. and Teerenstra, S. (2015). *Power Analysis of Trials with Multilevel Data*. CRC Press.

NIH (2022). Parallel Group- or Cluster-Randomized Trials. Retrieved May 26, 2023. https://www.researchmethodsresources.nih.gov/methods/grt.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649.

Robitzsch, A. and Grund, S. (2021). *miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'*. R package version 3.11-6.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Sadeghi, B., Kaiser, L. L., Hanbury, M. M., Tseregounis, I. E., Shaikh, U., Gomez-Camacho, R., Cheung, R. C., Aguilera, A. L., Whent, L., and De La Torre, A. (2019). A three-year multifaceted intervention to prevent obesity in children of Mexican-heritage. *BMC Public Health*, 19(1):1–12.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

Siddique, J., Harel, O., and Crespi, C. M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *The Annals of Applied Statistics*, 6(4):1814.

Snijders, T. A. and Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage.

Staudt, A., Freyer-Adam, J., Ittermann, T., Meyer, C., Bischof, G., John, U., and Baumann, S. (2022). Sensitivity analyses for data missing at random versus missing not at random using latent growth modelling: a practical guide for randomised controlled trials. *BMC Medical Research Methodology*, 22(1):250.

van Buuren, S. (2011). Multiple imputation of multilevel data. In Hox, J. J. and Roberts, J. K., editors, *Handbook of Advanced Multilevel Analysis*, pages 173–196. Routledge, New York, NY.

van Buuren, S. (2018). *Flexible Imputation of Missing Data.* CRC Press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67.

van der Leeden, R. and Busing, F. (1994). *First iteration versus IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3.* Unpublished manuscript. Department of Psychometrics and Research Methodology, Leiden University.

van der Leeden, R., Busing, F., and Meijer, E. (1997). *Bootstrap methods for two-level models.* Paper presented at the Multilevel Conference. Amsterdam.

White, I. R. (2010). simsum: Analyses of simulation studies including monte carlo error. *The Stata Journal*, 10(3):369–385.

White, I. R., Horton, N. J., Carpenter, J., Pocock, S. J., et al. (2011a). Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ*, 342.

White, I. R., Royston, P., and Wood, A. M. (2011b). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

Yan, X., Lee, S., and Li, N. (2009). Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):1085–1098.