A pattern-mixture model approach for handling missing continuous outcome data in longitudinal cluster randomized trials

Mallorie H. Fiero^{1§}, Chiu-Hsieh Hsu^{2§}, Melanie L. Bell^{2§}

¹Office of Biostatistics, Center for Drug Evaluation and Research,
U.S. Food & Drug Administration, Silver Spring, Maryland 20993

²Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman
College of Public Health, University of Arizona, Tucson, Arizona 85724

[§]Corresponding author

Contact detail:

10903 New Hampshire Ave., Silver Spring, Maryland 20993
1 (301) 348-1569

Email addresses:

MHF: Mallorie.Fiero@fda.hhs.gov

CH: pchhsu@email.arizona.edu

MLB: melaniebell@email.arizona.edu

Word count:

Disclaimer: This article reflects the views of the author and should not be construed to represent FDA's views or policies. This work was completed at the University of Arizona.

Summary

We extend the pattern mixture approach to handle missing continuous outcome data in longitudinal cluster randomized trials, which randomize groups of individuals to treatment arms, rather than the individuals themselves. Individuals who drop out at the same time point are grouped into the same dropout pattern. We approach extrapolation of the pattern mixture model by applying multilevel multiple imputation, which imputes missing values while appropriately accounting for the hierarchical data structure found in cluster randomized trials. To assess parameters of interest under various missing data assumptions, imputed values are multiplied by a sensitivity parameter, k , which increases or decreases imputed values. Using simulated data, we show that estimates of parameters of interest can vary widely under differing missing data assumptions. We carry out a sensitivity analysis using real data from a cluster randomized trial by increasing k until the treatment effect inference changes. By performing a sensitivity analysis for missing data, researchers can assess whether certain missing data assumptions are reasonable for their cluster randomized trial.

KEY WORDS: cluster randomized trials; missing data; pattern mixture model; multiple imputation

1 Introduction

1.1 Cluster randomized trials

Cluster randomized trials (CRTs), which randomly allocate groups of individuals to treatment arms rather than the individuals themselves, are becoming increasingly popular in health research [1]. This design is often chosen to minimize treatment arm contamination or to enhance compliance among participants. In CRTs, we cannot assume independence among individuals within the same cluster because of their similarity, which leads to decreased statistical power compared to individually randomized trials. The intraclass correlation coefficient (ICC), or ρ , is crucial in the design and analysis of CRTs, and measures the proportion of total variance due to clustering. The ICC ranges from 0-1 with 0 indicating responses within a cluster are independent, and 1 indicating responses within a cluster are all the same. Ignoring clusters in statistical analysis can lead to falsely low p-values, shortened confidence intervals, and an increased risk of obtaining significant results when there are none [2].

1.2 Missing data

Missing data are common in clinical trials and can lead to a reduction of power and bias in some cases. The missing data mechanism is the underlying reason why the data are missing. Missing data are said to be missing completely at random (MCAR) if the reason for a missing observation is unrelated to values of the outcome and covariates. However, MCAR is a very strong assumption and unlikely in clinical trials. A more reasonable assumption is missing at random (MAR), which requires that missingness is independent of the pattern of missing values after conditioning on fully observed values. Missing data are considered missing not at random (MNAR) if missingness depends on the unseen value of that observation after conditioning on fully observed data [3]. When data are MNAR, observations for those who drop out cannot be reliably predicted using observed data since the distribution differs between observed and missing observations [4]. For this reason, modeling dropout might be necessary in order to obtain correct inferences [5].

We consider continuous missing data at the individual level, though missing data can also occur at the cluster level in CRTs (for example, entire clusters missing). The likelihood of missingness in CRTs can depend on both cluster and individual level features, both of which can be used to recover information for missing data. We focus our attention on monotone missing data, in which individuals are observed until they drop out and their data from that time point until the end of the study is unobserved.

1.3 Sensitivity analysis for missing data

A sensitivity analysis for missing data is important in CRTs, as it evaluates the robustness of results based on differing missing data assumptions. The sensitivity analysis for missing data should be pre-specified in the trial protocol, and should include all individuals randomized. The primary analysis should be performed under the most plausible assumption, such as MAR, with a sensitivity analysis examining results based on departures from this assumption [6].

It has been suggested that researchers weaken the missing data assumption from the primary analysis [4]. In particular, researchers should carry out the primary analysis under MAR and sensitivity analysis under MNAR, as it is not possible to distinguish between MAR and MNAR data since the data are missing by definition. If results do not substantially change under departures from MAR, then the analysis is said to be robust. Despite these recommendations, a recent review evaluating handling of missing data in CRTs [7] found that 14 (16%) of the 86 reviewed trials reported performing a sensitivity analysis for missing data, with only five of them weakening the missingness assumption from the primary analysis. Three used multiple imputation, which takes into account uncertainty by replacing missing values with a set of possible values, and two used a likelihood based mixed model. Both methods are valid (produces unbiased estimates) under the MAR assumption. None of the trials included in the systematic review reported using MNAR models. Although strategies to deal with missing data in CRTs have been considered by some [8–11], none have developed methods to handle MNAR data. For this reason, we present a pattern mixture approach to handle MNAR data within the context of CRTs.

1.4 MNAR models

Two main approaches that have been proposed to handle longitudinal MNAR data include selection models [12] and pattern mixture models (PMMs) [13, 14]. These differ in the way the joint-distribution of the outcome and missing data mechanism are factorized. Selection models specify the joint distribution through the marginal distribution of the measurements and the conditional distribution of the missing data given the measurements. However, selection models are highly sensitive to specification of the measurement and dropout model, and require strong assumptions to describe the potential dropout patterns. This has led to PMMs receiving increased attention [5, 15]. PMMs specify the joint distribution through the marginal distribution of the missing data and the conditional distribution of the measurements given missing data. Individuals are grouped based on time of dropout. For example, in the simplest CRT scenario of two time points (baseline and follow-up) and assuming all individuals were measured at baseline, there are two possible dropout patterns: (1) responders - individuals who were measured at both baseline and follow-up, and (2)

non-responders - individuals who were measured at baseline, but not at follow-up. The individuals who drop out are assumed to have a different clinical outcome than the observed outcomes of those who remain in the trial. PMMs may be more easily understood by applied researchers and clinicians working on clinical trials because the observed data distribution and prediction distribution of missing data are explicitly separated [4, 12].

A critical issue of PMMs is that they are under-identified, which means that some parameters cannot be directly estimated because the non-responder dropout group does not have enough information to derive the distribution of the unobserved responses. Additional assumptions must be made to estimate all parameters in the non-responder dropout pattern. Nevertheless, some have argued that the under-identification issue is a benefit because it forces the researcher to think about the assumptions being made about the data [5, 15]. There are several techniques that have been proposed to deal with under-identification [16]. For example, Little proposed identifying restrictions, which link the inestimable parameters to parameters of the observed data model [13–15]. In a longitudinal trial with several timepoints, the large number of dropout patterns can be collapsed for simplification. Although this method is simple, there are strong untestable assumptions being made when grouping dropout patterns.

Another approach to overcome under-identification is to incorporate multiple imputation (MI), which takes uncertainty into account by imputing each missing value with a set of possible values under the MAR assumption. An imputation model is specified using observed data to estimate multiple values and create m complete datasets. Each completed dataset is then analyzed using standard statistical techniques and combined for inference [17]. When performing MI in longitudinal trials, the data are reshaped to wide form so that each row contains all measurements for each individual and relationships between measurements are preserved during imputation. With the added cluster level found in CRTs, standard MI leads to underestimated standard errors and confidence intervals that are too narrow since clusters are ignored [8].

In order to appropriately account for the multilevel structure of CRTs, the cluster feature needs to be incorporated into PMMs. Thus, we approach the under-identification problem of PMMs in the CRT context by applying multilevel MI, which accounts for the clustered structure and estimates appropriate standard errors [18]. We multiply MAR imputed values of the non-responders by a sensitivity parameter k to create MNAR imputed values in order to evaluate results under differing missing data assumptions.

1.5 Objectives

In Section 2, we provide an overview of the pattern mixture approach within the context of CRTs and describe multilevel MI. Sections 3 and 4 present a simulation study and appli-

cation to a dataset from the Postnatal Depression Economic Evaluation and Randomised Controlled Trial (PoNDER) study. Section 5 concludes with a discussion.

2 Pattern mixture models in cluster randomized trials

2.1 Linear mixed effects model

We focus on the simplest case of longitudinal data with two time points, where all individuals are observed at baseline and a proportion drop out at the follow-up time point, although this can be generalized to more than two time points. Consider a CRT with $i = 1, \dots, N$ clusters, $j = 1, \dots, n_i$ individuals per cluster, and $k = 1, \dots, t_{ij}$ measurements per individual. Let $Time_{ijk}$ denote the time of the k th measurement of individual j in cluster i , where $Time_{ijk} = 0$ denotes measurement at baseline and $Time_{ijk} = 1$ denotes measurement at follow-up. Further, suppose clusters were randomly allocated to the control arm, denoted as $Trt_i = 0$, or the treatment arm, denoted as $Trt_i = 1$. Consider the following mixed effects linear regression model with a single outcome of interest y_{ijk} :

$$y_{ijk} = \beta_0 + \beta_1(Time_{ijk}) + \beta_2(Trt_i) + \beta_3(Trt_i \times Time_{ijk}) + \gamma_i + \nu_{ij} + \epsilon_{ijk}, \quad (1)$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$ is the random effect at the cluster level and represents deviation of each cluster from the grand mean, $\nu_{ij} \sim N(0, \Sigma_\nu)$ is the random effect at the individual level and represents deviation of each individual from the cluster effect, and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ are the measurement errors terms. Furthermore, γ_i , ν_{ij} , and ϵ_{ijk} are assumed to be uncorrelated. The regression coefficient β_0 is the average response for the control arm ($Trt_i = 0$) at baseline ($Time_{ijk} = 0$), β_1 is the average difference in response between follow-up ($Time_{ijk} = 1$) and baseline ($Time_{ijk} = 0$) among individuals in the control arm ($Trt_i = 0$), β_2 is the average difference in response between the treatment arm ($Trt_i = 1$) and control arm ($Trt_i = 0$) at baseline ($Time_{ijk} = 0$), and β_3 is the average difference in slopes between treatment arms.

Let N_i denote the total number of measurements in cluster i where $N_i = \sum_{j=1}^{n_i} t_{ij}$. Generally, the mixed model with a single random cluster effect can be written as follows

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\nu}_i + \boldsymbol{\epsilon}_i, \quad (2)$$

where $\mathbf{Y}_i$ is an $N_i \times 1$ vector of responses, $\mathbf{X}_i$ is a known $N_i \times p$ design matrix of fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effects, and $\mathbf{Z}_i$ is a known $N_i \times u$ design matrix of random effects. Furthermore, $\boldsymbol{\nu}_i$ is a $u \times 1$ vector of unknown random effects distributed $N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\epsilon}_i$ is an $N_i \times 1$ vector of random residuals distributed $N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{N_i})$, where $\mathbf{I}_{N_i}$ represents the $N_i \times N_i$ identity matrix.

2.2 Pattern mixture models

Little proposed PMMs for repeated measures with dropouts where the MAR assumption is too strong [13]. Let $\mathbf{R}$ be the vector of missingness indicators for the response vector $\mathbf{Y}$ with $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ denoting observed and unobserved responses, respectively. Further, let $\mathbf{X}$ be a set of observed covariates. Pattern mixture models (PMMs) factorize the joint-distribution of the response and missing data mechanism by:

$$p(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \mathbf{R}|\mathbf{X}) = p(\mathbf{R}|\mathbf{X})p(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}|\mathbf{R}, \mathbf{X}), \quad (3)$$

where $p(\mathbf{R}|\mathbf{X})$ is the conditional probability distribution of the dropout pattern given observed covariates and $p(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}|\mathbf{R}, \mathbf{X})$ is the probability distribution of the response vector given the dropout pattern and observed covariates [13].

2.3 Transforming MAR imputed values to create MNAR imputed values

Rubin and Little have both advocated for the use of simple techniques such as multiplying imputed values by a factor, as they are transparent, readily understandable, and can be easily implemented in current statistical software [17, 19, 20]. Thus, we employ multilevel MI (described below) and multiply MAR imputed values by a sensitivity parameter k to generate MNAR imputed values such that [17]

$$(\text{MNAR imputed } Y_i) = k \times (\text{MAR imputed } Y_i). \quad (4)$$

For example, if $k = 1.3$ or $k = 0.8$, MAR imputed values are increased by 30% or decreased by 20%, respectively. This creates MNAR observations because the missing data of the non-responders are systematically higher or lower than the observed data of the responders. In the case that the MAR imputed value is negative, a more general version of Equation 4 is

$$(\text{MNAR imputed } Y_i) = [(k - 1) \times |\text{MAR imputed } Y_i|] + \text{MAR imputed } Y_i, \quad (5)$$

where negative imputed values are increased when $k > 1$ and decreased when $k < 1$. When multiplying MAR imputed values by a factor of k , imputations should be checked to identify that the MNAR imputed values fall within a realistic range of the data.

2.4 Multilevel multiple imputation

Multilevel MI applies the Gibbs sampler to impute missing data found in hierarchical data. The Gibbs sampler is a Markov chain Monte Carlo (MCMC) sampling technique

for sampling from multivariate probability distributions [21, 22]. Using the linear mixed model given in Equation 2, multilevel MI simulates the distribution of parameters using MCMC methods with the following steps:

1. Sample β from $p(\beta|\mathbf{y}, \boldsymbol{\nu}, \sigma^2)$
2. Sample $\boldsymbol{\nu}$ from $p(\boldsymbol{\nu}|\mathbf{y}, \beta, \boldsymbol{\Sigma}, \sigma^2)$
3. Sample $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma}|\boldsymbol{\nu})$
4. Sample σ^2 from $p(\sigma^2|\mathbf{y}, \beta, \boldsymbol{\nu})$
5. Repeat steps 1-4 until convergence
6. Sample $\mathbf{y}_{mis}$ from $p(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \beta, \boldsymbol{\nu}, \boldsymbol{\Sigma}, \sigma^2)$

where $\mathbf{y}$ represents the response vector, with $\mathbf{y}_{obs}$ and $\mathbf{y}_{mis}$ denoting observed and unobserved responses, respectively. Under the MAR assumption, the parameter distribution is simulated in steps 1-5 using observed data such that $\mathbf{y}$ is replaced by $\mathbf{y}_{obs}$. Imputations for missing data are created in step 6 and are calculated by drawing from

$$\begin{aligned}\boldsymbol{\epsilon}_i^* &\sim N(0, \sigma^2) \\ \mathbf{y}_i^* &= \mathbf{X}_i\beta + \mathbf{Z}_i\boldsymbol{\nu}_i + \boldsymbol{\epsilon}_i^*\end{aligned}$$

where the parameters on the right side of the equation are replaced by values drawn under the Gibbs sampler described above [18].

2.5 Combining inferences

Once the imputations are generated, the m completed datasets are analyzed without accounting for dropout in the analysis model. The point estimate and corresponding standard error for a parameter of interest Q are combined for inference using Rubin's Rules, which account for within and between imputation variability [17]. Let $\hat{Q}_l$ and $\hat{W}_l$ be the point and variance estimates, respectively, obtained from $l = 1, \dots, m$ imputed datasets. The overall point estimate for Q is the mean over the imputed datasets:

$$\bar{Q} = \frac{1}{m} \sum_{l=1}^m \hat{Q}_l.$$

The overall standard error is $\sqrt{T}$,

$$T = \bar{W} + \left(1 + \frac{1}{m}\right) B,$$

where $\bar{W} = \frac{1}{m} \sum_{l=1}^m \hat{W}_l$ is the within-imputation variance and $B = \frac{1}{m-1} \sum_{l=1}^m (\hat{Q}_l - \bar{Q})^2$ is the between-imputation variance. Confidence intervals and tests are approximated with $(Q - \bar{Q})/\sqrt{T} \sim t_{v^*}$ with adjusted degrees of freedom computed as

$$v^* = \left[\frac{1}{v} + \frac{1}{\hat{v}_{obs}} \right],$$

where $v = (m-1)(1+r^{-1})^2$, $r = (1+m^{-1})B/\bar{W}$, $\hat{v}_{obs} = (1-\gamma)v_0(v_0+1)/(v_0+3)$, $\gamma = (1+m^{-1})B/T$, and v_0 indicates the degrees of freedom for the complete data [23].

3 Simulation study

3.1 Data generation

We based our simulation study on our application example in Section 4 when generating data and choosing values for the sensitivity parameter k . We mimicked the scenario in which the treatment arm had different non-responders compared to the responders, while the control arm had similar non-responders compared to the responders. Adding to Equation 1, a CRT with two time points $(y_1, y_2)^T$ and missing data at the follow-up time point was simulated under the following clustered pattern-mixture model [24]:

$$\begin{aligned} y_{ijk} = & \beta_0 + \beta_1(Time_{ijk}) + \beta_2(Trt_i) + \beta_3(Trt_i \times Time_{ijk}) \\ & + \beta_4(Drop_{ijk} \times Trt_i \times Time_{ijk}) + \gamma_i + \nu_{ij} + \epsilon_{ijk} \end{aligned} \quad (6)$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$ denotes the random cluster effect, $\nu_{ij} \sim N(0, \Sigma_\nu)$ denotes the random individual effect, and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ denotes the measurement errors terms. $Time_{ijk}$ was coded as 0 for baseline and 1 for follow-up, Trt_i was coded as 0 for control and 1 for treatment, and $Drop_{ijk}$ was coded as 0 for responder and 1 for non-responder at follow-up. We consider a CRT where a higher y_{ijk} value indicates a worse outcome, such as depression, as in our application example in Section 4. The regression coefficients were defined as $\beta_0 = 7$, $\beta_1 = -1$, $\beta_2 = 0$, $\beta_3 = -2$, $\beta_4 = 3$. This simulates a CRT in which there is no difference in the mean response between the treatment arms at baseline, and a lower mean response for the treatment arm compared to the control arm at follow-up. The random individual effect ν_{ij} and residuals ϵ_{ijk} were both normally distributed with a mean of 0 and variance of 12. We varied ρ from 0.001 to 0.5. In practice, the ICC is rarely above 0.1, but we included higher values of ICC to assess behavior under extreme cases. The total number of clusters and cluster size varied as (12, 15), (12, 30), (12, 100), (30, 15), (30, 30), (30, 100). We allocated an equal number of clusters to each treatment arm.

We simulated a 40% dropout rate at follow-up, which means that 40% of individuals in each treatment arm had a value of 1 for $Drop_{ijk}$ at follow-up and were deleted. The sensitivity

parameter is β_4 , which computes to $k = 1.0$ for the control arm and $k = 1.75$ for the treatment arm. This increases the mean of the non-responders in the treatment group and creates MNAR data in the treatment arm when the data are deleted. For the control arm, the mean response of the unobserved data of the non-responders is the same as the mean response of the observed data of the responders at follow up. For the treatment arm, the mean response of the unobserved data of the non-responders is 75% higher than the mean response of the observed data of the responders at follow-up. The variance of y_2 for the non-responders is assumed to be the same as the variance of y_2 for the responders.

3.2 Methods

We drew 500 samples from each scenario and carried out multilevel MI to impute missing y_2 values using the `mice` package in R version 3.2.3 [25]. We carried out multilevel MI for each treatment arm separately ($m = 5$ imputation sets), and included y_1 in the imputation model.

For the control arm, we multiplied the imputed values by $k = 1.0$, which assumes that the unobserved outcomes of those who dropped out in the control arm are similar to the observed outcomes of those who remained in the trial (MAR). For the treatment arm, we multiplied each imputed value by $k = (0.8, 1.0, 1.3, 1.7)$. These k values represent a range of differing clinical assumptions regarding the unobserved data in the treatment arm. A multiplier of $k = 1.0$ assumes that the unobserved data are similar to the observed data (MAR). A multiplier of $k = 1.3$ increases the imputed values by 30%, and assumes that the unobserved outcomes are slightly higher than the observed outcomes. For example, individuals who dropped out had a slightly poorer outcome than the individuals remained in the trial. A multiplier of $k = 1.7$ increases the imputed values by 70%, and assumes that the unobserved outcomes are much higher (i.e., much poorer outcome) than the observed outcomes. A multiplier of $k = 0.8$ decreases the imputed values by 20%, and incorrectly assumes that the unobserved outcomes are lower (i.e., better outcome) than the observed outcomes [24].

Using the completed dataset, we modeled the outcome with a mixed model (`lme4` package in R) using Equation 6, but without including the $Drop_{ijk}$ term. The following parameters of interest were calculated: (1) change over time in the treatment arm and (2) treatment effect, defined as the mean difference in arms at follow-up. Their corresponding standard errors were also calculated. Using the regression coefficients in Equation 6, the true change over time in the treatment arm was,

$$\begin{aligned} & \{[0.60 \times (\beta_0 + \beta_1 + \beta_2 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4)]\} \\ & - \{[0.60 \times (\beta_0 + \beta_2)] + [0.40 \times (\beta_0 + \beta_2)]\} = -1.8. \end{aligned}$$

and the true treatment effect was,

$$\begin{aligned} & \{[0.60 \times (\beta_0 + \beta_1 + \beta_2 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4)]\} \\ & - \{[0.60 \times (\beta_0 + \beta_1)] + [0.40 \times (\beta_0 + \beta_1)]\} = -0.8. \end{aligned}$$

Parameter estimates were pooled using Rubin’s rules as implemented in `mice` [17]. For both parameters of interest, we computed the following measures of performance:

1. Percent bias: the difference between the true value and estimate of the fixed parameter, divided by the true value
2. Coverage: proportion of times the true value was contained in the 95% confidence interval of the fixed parameter estimates, change over time in the treatment arm and treatment effect
3. Ratio of model-based to empirical standard error

3.3 Results

We present the results of our simulations in Tables 1 - 3. Table 1 displays the percent bias of the treatment arm change over time and treatment effect under each sensitivity parameter k . As expected, the percent bias for both estimates is smallest for $k = 1.7$, as it is closest to the true sensitivity parameter. Percent bias for change over time in the treatment arm and treatment effect are -3.0% and -9.3%, respectively, for 12 total clusters, 30 individuals per cluster, and an ICC of 0.01. Under the MAR assumption ($k = 1.0$) and the incorrect MNAR assumption ($k = 0.8$), the estimates have a severe downward bias. Under the same scenario, percent bias for change over time in the treatment arm for $k = 1.0$ and $k = 0.8$ are -149.7% and -189.8%, respectively. Percent bias is more extreme in the treatment effect. For example, the percent bias for $k = 1.3$ is -38.7% for change over time in the treatment arm, and -89.5% for the treatment effect.

Table 2 presents the coverage of nominal 95% confidence intervals for both estimates. Coverage of the treatment arm change over time and treatment effect estimates increase as k becomes closer to the true sensitivity parameter, and is highest for $k = 1.7$. For example, with 12 total clusters, 100 individuals per cluster, and an ICC of 0.1, the coverage for the treatment effect estimate was 92.4% for $k = 1.7$, and 75.4% for $k = 1.0$. Furthermore, coverage for both estimates decreases as ICC increases as seen under $k = 1.7$. For 12 clusters and 100 individuals per cluster, coverage of the treatment effect under $k = 1.7$ was highest for an ICC of 0.001 (96.4%) and lowest for an ICC of 0.5 (91.6%), though the PMM still performed fairly well under extreme ICC.

Table 3 displays the ratios of model-based to empirical standard errors for change over time in the treatment arm and treatment effect. Model-based standard errors are expected

to increase with respect to empirical standard errors since multilevel MI incorporates uncertainty due to missing data. Overall, results were similar for the percent bias of the treatment arm change over time and treatment effect. More extreme values of k overestimates the standard errors because the imputed values are multiplied, which increases variances of the estimates. For 30 total clusters, 30 individuals per cluster and an ICC of 0.3, the ratio of model-based to the empirical standard error for the treatment effect was 1.06 under $k = 1.0$ and 1.14 under $k = 1.7$. The results did not change substantially when the cluster size was smaller. For a similar scenario of 30 total clusters and an ICC of 0.3, the ratio of model-based to the empirical standard error for the treatment effect was 1.01 under $k = 1.0$ and 1.12 under $k = 1.7$ when the number of individuals per cluster decreased to 15.

4 Application to the PoNDER study

4.1 The data

The Postnatal Depression Economic Evaluation and Randomised Controlled Trial (PoNDER) study assessed whether training health visitors (HV) to provide psychologically informed sessions improved depressive symptoms among postnatal women. This study has been described elsewhere [26]. Briefly, general practitioner (GP) practices were randomized to HV training (treatment) or HV usual care (control). There were a total of 37 ($N = 1,151$) and 63 ($N = 2,268$) GP practices in the control and treatment arm, respectively. The average number of individuals per cluster was 34 (range 1-119). Depression among postnatal women was measured using the 10-item Edinburgh Postnatal Depression Scale (EPDS), which ranges from 0-30 with higher scores indicating worse outcomes. Measurements were scheduled at baseline and 6 months.

We included all participants who were observed at baseline. Table 4 displays the means and standard deviations of EPDS score by treatment arm at baseline. For the control and treatment arms, 237 (20.6%) and 523 (23.1%) dropped out at the 6-month follow-up, respectively. For the treatment arm, those who dropped out had a higher average EPDS score at baseline (mean (standard deviation (SD)) = 8.0 (5.9)) compared to those who did not drop out (mean (SD) = 6.6 (4.8)). This shows the importance of analyzing the responders and non-responders separately in a sensitivity analysis for missing data to evaluate how results change under differing missingness assumptions.

4.2 Methods

Since the baseline EPDS score for the non-responders were similar to the responders in the control arm, we carried out a sensitivity analysis assuming MAR for the non-responders in the control arm and MNAR for the non-responders in the treatment arm. For each treatment arm, we carried out a multilevel MI ($m = 5$) with baseline EPDS score included as a covariate in the imputation model. For the treatment arm, we increased the sensitivity parameter by increments of 10%, indicating a worsening of the outcome for the non-responders (i.e., 1.0, 1.1, 1.2, etc). We continued to increase the sensitivity parameter until the treatment effect inference changed. Figure 1 graphically displays the trajectory of the non-responders under $k = 1.0$ for the control arm and varying k for the treatment arm.

For each multiply imputed dataset, we carried out a mixed model adjusting for GP practices and individuals as random effects, and computed the (1) change over time for the treatment arm and (2) treatment effect, defined as the mean difference in arms post-treatment ($y_{2T} - y_{2C}$). Inferences were combined using Rubin’s rules.

4.3 Results

Table 5 displays the results of each PMM scenario. As k increases the slope of the treatment arm as well as the treatment effect attenuate. For example, the change in treatment arm over time was estimated at -1.44 (95% CI = -1.69, -1.19) under the MAR assumption, and -0.82 (95% CI = -1.11, -0.53) under $k = 1.5$. The inference of the change in EPDS score for the treatment arm remained similar to the MAR assumption. The inference of the treatment effect changed at $k = 1.5$ (Treatment effect = -0.36, 95% CI = -0.85, 0.13), which assumes that the non-responders in the treatment arm had a worse EPDS score by 50%. At this point, researchers can evaluate whether this assumption is reasonable and report results for this range of k as their sensitivity analysis for missing data. The ICC remained at 0.01 for all PMM scenarios.

5 Discussion

Missing data are prevalent in CRTs. It is crucial to handle missing data with appropriate methods in order to increase statistical power and reduce the possibility of bias in estimating the treatment effect. Despite recommendations to carry out a sensitivity analysis for missing data [4], very few CRTs have reported performing a sensitivity analysis in practice [7]. To facilitate performing sensitivity analyses for missing data in CRTs, we have proposed an approach within the pattern mixture framework to analyze clustered MNAR

data. We implemented multilevel MI in order to account for the clustered data structure of CRTs, then multiplied MAR imputed values by a factor, k to increase or decrease imputed values and create MNAR imputed values.

Multilevel MI should be used when imputing missing data in CRTs because it incorporates the multilevel data structure and produces appropriate standard errors for estimates of interest. Van Buuren showed that ignoring clustering in MI produces severely biased variance components when the data are clustered ($ICC > 0$) [18]. Despite recommendations from statisticians to incorporate clusters into imputation methods [8, 11, 18], none of the trials that implemented MI accounted for clustering in the recent systematic review evaluating handling of missing data in CRTs [7]. Multilevel MI can be implemented using the `mice` package in R, which can impute missing individual level outcomes and covariates. Mistler provides a SAS macro for implementing multilevel MI called `MMI_IMPUTE`, which can impute both individual and cluster level variables [27].

Standard errors are subject to over-inflation when multiplying imputed values by k , especially with extreme values of k . Transformed MNAR values should be checked to ensure imputations lie within an appropriate range of the data. Another simple approach is to carry out multilevel MI and add or subtract imputed values by δ , where δ is the mean difference in the outcome between the responders and non-responders [28]. This shifts the imputed values of the non-responders, while preserving the standard errors of the estimates of interest. Choosing the value of k or δ heavily depends on the subject matter of the trial, and should be elicited from experts in the field, such as the trial investigators or experts not committed to the trial. For example, White and colleagues collected opinions of several experts using a questionnaire to obtain information about plausible differences between responders and non-responders [29]. A range of plausible k or δ can be specified, or an average can be specified if a single analysis is preferred.

We used our application example as motivation in our simulation study where we assumed the non-responders were different from the responders in the treatment arm, while the non-responders were similar to the responders in the control arm. In reality, non-responders may be different compared to the responders in both treatment arms. In this case, an appropriate range of k can be used for both treatment arms in a sensitivity analysis for missing data. Different ranges of k can be based on the treatment arm, reason for missingness, or time of dropout. For example, individuals who were lost to follow-up can be assumed MAR dropout, while individuals who withdrew could be considered MNAR dropout.

We considered the simplest case of longitudinal data with two time points and two missing data patterns: responders and non-responders. In practice, there can be multiple post-baseline measurements in longitudinal CRTs, which makes the PMM more complex due to the increased number of missing data patterns. One approach to generalize the PMM in CRTs to more than two time points is to initially perform multilevel MI to impute missing outcome observations under the MAR assumption. In order to transform MAR imputed

values into MNAR imputed values, the multiplier k can be specified at each time point [30] or can be specified at the first missed response and then decreased by a certain fraction with every missed response. Longitudinal trials with more than two time points should be further investigated within the CRT context.

Through our simulation study, we showed that estimates of parameters of interest can greatly differ depending on the missing data assumption. For this reason, it is important to carry out a sensitivity analysis to assess the robustness of the primary results under differing missing data assumptions, as we did with the PoNDER study. The treatment effect inference attenuated with higher values of k , and changed when the imputed EPDS scores of the non-responders were increased by 50%. By doing this, researchers can examine the impact of departure from the MAR assumption.

Other approaches for MNAR missing data that have been proposed, but not yet investigated within the CRT scenario include identifying restrictions [15, 16], selection models [12], and MNAR approximate Bayesian Bootstrap [31, 32]. Consideration of models depends on plausible assumptions of the missing data for the particular trial, as well as ease of interpretation for trial investigators.

References

1. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice* 2000; **17**(2):192–196.
2. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978; **108**(2):100–102.
3. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**(3):581–592.
4. Council NR. *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press: Washington DC, 2010.
5. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.
6. White IR, Horton NJ, Carpenter J, Pocock SJ, *et al.*. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2011; **342**:d40.
7. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomised trials: a systematic review. *Trials* 2015; **17**:72.
8. Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal* 2008; **50**(3):329–345.

9. Andridge RR. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal* 2011; **53**(1):57–74.
10. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology* 2011; **11**(1):1.
11. Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat* 2012; **2**:93–103.
12. Little RJ, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, 2014.
13. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**(421):125–134.
14. Little RJ. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; **81**(3):471–483.
15. Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. *Biostatistics* 2002; **3**(2):245–265.
16. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2003; **22**(16):2553–2575.
17. Rubin DB. *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 1987.
18. Van Buuren S, et al.. *Multiple imputation of multilevel data*. Routledge New York, NY, 2011.
19. Little RJ. Comments on: Missing data methods in longitudinal studies: a review. *Test* 2009; **18**(1):47–50.
20. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software* 2011; **45**(3).
21. Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 1984; **6**:721–741.
22. Casella G, George EI. Explaining the gibbs sampler. *The American Statistician* 1992; **46**(3):167–174.
23. Barnard J, Rubin DB. Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika* 1999; **86**(4):948–955.

24. Siddique J, Harel O, Crespi CM. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *The Annals of Applied Statistics* 2012; **6**(4):1814.
25. Van Buuren S, Oudshoorn C. mice: Multivariate imputation by chained equations. r package version 1.16 2007.
26. Morrell CJ, Warner R, Slade P, Dixon S, Walters S, Paley G, Brugha T. *Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation: the PoNDER trial*. Prepress Projects, 2009.
27. Mistler SA. A sas macro for applying multiple imputation to multilevel data. *Proceedings of the SAS Global Forum*, 2013.
28. Van Buuren S, Boshuizen HC, Knook DL, *et al.*. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**(6):681–694.
29. White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials* 2007; **4**(2):125–139.
30. Carpenter J, Kenward M. *Multiple imputation and its application*. John Wiley & Sons, 2012.
31. Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* 1991; **10**(4):585–598.
32. Siddique J, Belin TR. Using an approximate bayesian bootstrap to multiply impute nonignorable missing data. *Computational Statistics & Data Analysis* 2008; **53**(2):405–415.

Table 1: Percent bias of change over time in the treatment arm and treatment effect with MNAR data in y_{ijk} .

No. clusters	Cluster size	ICC	Treatment arm change				Treatment effect ¹			
			k				k			
			0.8	1.0	1.3	1.7	0.8	1.0	1.3	1.7
12	15	0.001	-82.6	-65.0	-38.5	-2.4	-195.8	-156.2	-96.5	-15.4
		0.01	-86.2	-69.0	-41.2	-5.6	-184.6	-145.8	-83.4	-3.2
		0.1	-80.4	-61.9	-34.5	1.4	-185.2	-143.6	-81.9	-1.1
		0.3	-85.7	-67.8	-41.7	-6.6	-188.2	-147.7	-89.2	-10.1
		0.5	-85.4	-67.2	-40.9	-5.8	-186.6	-145.6	-86.4	-7.4
12	30	0.001	-82.8	-64.9	-38.2	-2.5	-186.8	-146.7	-86.5	-6.3
		0.01	-83.2	-65.4	-38.7	-3.0	-189.8	-149.7	-89.5	-9.3
		0.1	-85.0	-67.5	-41.1	-6.0	-195.1	-155.5	-96.3	-17.2
		0.3	-84.8	-66.7	-39.6	-3.5	-180.7	-140.1	-79.1	2.2
		0.5	-84.4	-66.8	-40.3	-5.1	-182.3	-142.6	-83.1	-3.8
12	100	0.001	-83.8	-65.9	-39.1	-3.5	-186.7	-146.6	-86.4	-6.1
		0.01	-83.5	-65.6	-38.6	-2.7	-183.8	-143.4	-82.8	-2.0
		0.1	-84.5	-66.7	-40.1	-4.5	-185.9	-145.9	-85.8	-5.8
		0.3	-84.2	-66.5	-40.0	-4.6	-195.7	-156.0	-96.3	-16.8
		0.5	-85.2	-67.3	-40.3	-4.3	-189.6	-149.2	-88.5	-7.6
30	15	0.001	-84.4	-67.0	-40.5	-5.1	-190.2	-151.1	-91.3	-11.9
		0.01	-82.4	-64.1	-37.9	-3.1	-187.2	-146.0	-87.0	-8.8
		0.1	-84.3	-66.7	-39.7	-3.5	-184.4	-144.9	-84.2	-2.7
		0.3	-83.9	-66.3	-39.6	-4.0	-190.5	-150.9	-90.9	-10.7
		0.5	-84.6	-66.8	-40.1	-4.2	-184.1	-144.1	-84.0	-3.4
30	30	0.001	-84.0	-66.1	-39.3	-3.51	-188.8	-148.6	-88.2	-7.8
		0.01	-86.0	-68.3	-41.7	-6.3	-190.5	-150.6	-90.9	-11.2
		0.1	-84.4	-66.7	-40.1	-4.6	-192.7	-152.8	-92.9	-13.0
		0.3	-84.0	-66.6	-40.4	-5.6	-198.9	-159.7	-100.9	-22.5
		0.5	-84.7	-66.9	-40.1	-4.4	-204.2	-164.0	-103.8	-23.5
30	100	0.001	-83.9	-66.2	-39.6	-4.2	-190.2	-150.3	-90.5	-10.7
		0.01	-84.4	-66.6	-40.0	-4.5	-190.3	-150.4	-90.5	-10.7
		0.1	-84.9	-67.0	-40.1	-4.36	-184.4	-144.2	-83.8	-3.3
		0.3	-84.3	-66.3	-39.3	-3.34	-183.2	-142.7	-82.0	-1.1
		0.5	-83.6	-65.8	-39.0	-3.4	-190.4	-150.3	-90.1	-10.0

Abbreviations: MNAR, missing not at random; ICC, intraclass correlation coefficient

¹Treatment effect: mean difference between treatment arms at follow-up

Table 2: Coverage of nominal 95% confidence intervals of true values for change over time in the treatment arm and treatment effect.

No. clusters	Cluster size	ICC	Treatment arm change				Treatment effect ¹			
			k				k			
			0.8	1.0	1.3	1.7	0.8	1.0	1.3	1.7
12	15	0.001	37.6	68.2	92.6	98.0	64.2	77.6	91.8	98.0
		0.01	36.8	62.6	91.0	97.6	71.4	84.0	92.2	96.8
		0.1	40.0	67.8	92.0	96.2	77.8	86.8	92.8	95.4
		0.3	35.0	65.4	89.6	95.4	86.8	89.4	91.2	92.0
		0.5	36.4	64.0	89.4	95.8	87.8	89.4	91.0	91.2
12	30	0.001	11.2	41.8	84.6	97.6	46.6	67.0	88.4	97.8
		0.01	12.8	45.0	83.0	98.4	47.2	69.4	89.6	97.2
		0.1	8.6	37.2	81.4	97.2	66.4	76.2	89.0	92.4
		0.3	11.6	38.4	80.2	96.0	82.2	84.8	87.0	89.4
		0.5	12.8	37.4	79.0	92.2	88.8	89.2	89.8	88.6
12	100	0.001	0.4	7.8	52.8	97.6	5.6	24.2	67.0	96.4
		0.01	0.6	7.2	52.6	97.2	13.8	35.2	73.0	95.0
		0.1	1.0	6.4	54.0	96.2	65.6	75.4	88.0	92.4
		0.3	0.6	8.4	53.8	90.4	84.4	87.4	90.8	90.6
		0.5	0.4	6.8	54.0	83.8	88.2	89.0	90.6	91.6
30	15	0.001	5.8	29.8	77.0	97.0	31.8	58.4	85.0	96.4
		0.01	6.4	33.8	81.0	97.8	38.0	61.6	85.2	95.4
		0.1	9.0	30.6	77.2	97.4	56.2	71.4	89.2	95.2
		0.3	5.2	31.8	75.8	96.6	74.2	84.0	90.4	92.4
		0.5	5.6	32.0	77.8	94.4	86.0	89.2	91.2	91.8
30	30	0.001	1.0	12.4	63.2	97.0	9.4	31.4	74.2	95.6
		0.01	1.2	9.8	59.6	97.2	9.4	30.8	75.2	96.2
		0.1	0.6	10.2	62.8	96.0	38.6	57.0	82.4	92.2
		0.3	0.6	11.0	61.8	95.8	74.4	82.6	90.6	94.0
		0.5	0.6	12.4	62.8	94.0	83.0	86.2	89.2	92.0
30	100	0.001	0.2	2.4	28.2	97.4	1.8	4.6	38.0	95.4
		0.01	0.0	3.0	28.4	96.8	0.2	2.8	40.8	95.2
		0.1	0.2	3.2	31.0	95.0	33.2	53.6	78.6	92.4
		0.3	0.0	2.6	26.0	90.4	73.2	81.8	88.4	90.4
		0.5	0.2	4.4	28.2	82.6	85.2	88.2	91.8	92.6

Abbreviations: ICC, intraclass correlation coefficient

¹Treatment effect: mean difference between treatment arms at follow-up

Table 3: Ratios of model-based to empirical standard errors for change over time in the treatment arm and treatment effect.

No. clusters	Cluster size	ICC	Treatment arm change				Treatment effect ¹			
			k				k			
			0.8	1.0	1.3	1.7	0.8	1.0	1.3	1.7
12	15	0.001	1.177	1.280	1.445	1.736	1.230	1.284	1.386	1.562
		0.01	1.158	1.269	1.426	1.687	1.202	1.268	1.353	1.495
		0.1	1.092	1.193	1.352	1.629	1.104	1.139	1.205	1.311
		0.3	1.189	1.275	1.478	1.751	0.975	0.997	1.042	1.100
		0.5	1.240	1.332	1.535	1.870	0.956	0.976	1.009	1.060
12	30	0.001	1.169	1.265	1.444	1.722	1.296	1.353	1.460	1.629
		0.01	1.144	1.244	1.421	1.698	1.182	1.234	1.327	1.476
		0.1	1.153	1.246	1.419	1.690	0.983	1.011	1.058	1.131
		0.3	1.156	1.249	1.429	1.711	0.917	0.937	0.969	1.017
		0.5	1.094	1.179	1.348	1.620	0.953	0.972	1.002	1.045
12	100	0.001	1.090	1.180	1.341	1.595	1.191	1.243	1.337	1.487
		0.01	1.131	1.224	1.396	1.660	1.056	1.094	1.164	1.275
		0.1	1.150	1.244	1.422	1.693	1.012	1.036	1.072	1.127
		0.3	1.159	1.253	1.432	1.716	1.013	1.033	1.065	1.110
		0.5	1.221	1.316	1.506	1.816	0.991	1.011	1.042	1.084
30	15	0.001	1.145	1.227	1.390	1.661	1.228	1.283	1.377	1.554
		0.01	1.124	1.220	1.392	1.666	1.217	1.273	1.367	1.523
		0.1	1.175	1.252	1.436	1.706	1.112	1.146	1.211	1.311
		0.3	1.162	1.260	1.444	1.750	0.988	1.012	1.052	1.115
		0.5	1.144	1.241	1.418	1.730	0.944	0.965	0.996	1.045
30	30	0.001	1.196	1.294	1.477	1.758	1.213	1.270	1.372	1.530
		0.01	1.144	1.238	1.408	1.677	1.165	1.213	1.302	1.445
		0.1	1.140	1.230	1.401	1.670	1.016	1.043	1.091	1.163
		0.3	1.158	1.253	1.431	1.717	1.033	1.056	1.092	1.144
		0.5	1.154	1.248	1.428	1.727	0.947	0.967	0.997	1.041
30	100	0.001	1.147	1.241	1.413	1.678	1.262	1.317	1.421	1.586
		0.01	1.201	1.300	1.489	1.769	1.131	1.171	1.245	1.359
		0.1	1.139	1.232	1.410	1.681	0.993	1.015	1.051	1.103
		0.3	1.170	1.266	1.442	1.722	0.948	0.967	0.997	1.040
		0.5	1.163	1.250	1.432	1.725	0.983	1.002	1.034	1.077

Abbreviations: ICC, intraclass correlation coefficient

¹Treatment effect: mean difference between treatment arms at follow-up

Table 4: PoNDER study. Means and standard deviations of baseline EPDS score by treatment arm and dropout pattern.

Dropout pattern	Control N = 1151		Treatment N = 2268	
	N (%)	Mean (SD)	N (%)	Mean (SD)
Responders	914 (79.4)	6.8 (5.0)	1745 (76.9)	6.6 (4.8)
Non-responders	237 (20.6)	6.8 (5.1)	523 (23.1)	8.0 (5.9)

Abbreviations: EPDS, Edinburgh Postnatal Depression Scale; SD, standard deviation

Table 5: PoNDER study. Sensitivity analysis for missing data in 6-month EPDS score. Change in treatment arm over time and treatment effect results were assessed by increasing imputed values with a range of k .

k	Treatment arm change over time (95% CI)	p -value	Treatment effect ¹ (95% CI)	p -value
1.0	-1.44 (-1.69, -1.19)	<0.0001	-0.97 (-1.42, -0.52)	<0.0001
1.1	-1.31 (-1.57, -1.06)	<0.0001	-0.85 (-1.31, -0.40)	<0.001
1.2	-1.19 (-1.45, -0.93)	<0.0001	-0.73 (-1.19, -0.27)	0.002
1.3	-1.07 (-1.34, -0.79)	<0.0001	-0.61 (-1.08, -0.14)	0.012
1.4	-0.94 (-1.23, -0.66)	<0.0001	-0.48 (-0.96, -0.004)	0.048
1.5	-0.82 (-1.11, -0.53)	<0.0001	-0.36 (-0.85, 0.13)	0.146

Abbreviations: EPDS, Edinburgh Postnatal Depression Scale; CI, confidence interval

¹Treatment effect: mean difference between treatment arms at follow-up