STATISTICAL APPROACHES FOR HANDLING MISSING DATA IN CLUSTER RANDOMIZED TRIALS

by

Mallorie H. Fiero

A Dissertation Submitted to the Faculty of the

MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN BIOSTATISTICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Mallorie H. Fiero, titled *Statistical Approaches for Handling Missing Data in Cluster Randomized Trials* and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: 13 April 2016

Melanie Bell

_____ Date: 13 April 2016

Denise Roe

_____ Date: 13 April 2016

Chiu-Hsieh Hsu

_____ Date: 13 April 2016

Eyal Oren

Final approval and acceptance of this dissertation is contingent upon the candidates submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 13 April 2016

Dissertation Chair: Melanie Bell

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Mallorie H. Fiero

DEDICATION

To my husband and parents.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

In cluster randomized trials (CRTs), groups of participants are randomized as opposed to individual participants. 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 intracluster correlation coefficient (ICC) is crucial in the design and analysis of CRTs, and measures the proportion of total variance due to clustering. Missing data is a common problem in CRTs and should be accommodated with appropriate statistical techniques because they can compromise the advantages created by randomization and are a potential source of bias. In three papers, I investigate statistical approaches for handling missing data in CRTs.

In the first paper, I carry out a systematic review evaluating current practice of handling missing data in CRTs. The results show high rates of missing data in the majority of CRTs, yet handling of missing data remains suboptimal. Fourteen (16%) of the 86 reviewed trials reported carrying out a sensitivity analysis for missing data. Despite suggestions to weaken the missing data assumption from the primary analysis, only five of the trials weakened the assumption. None of the trials reported using missing not at random (MNAR) models.

Due to the low proportion of CRTs reporting an appropriate sensitivity analysis for missing data, the second paper aims to facilitate performing a sensitivity analysis for missing data in CRTs by extending the pattern mixture approach for missing clustered data under the MNAR assumption. I implement multilevel multiple imputation (MI) in order to account for the hierarchical structure found in CRTs, and multiply imputed values by a sensitivity parameter, k, to examine parameters of interest under different missing data assumptions. The simulation results show that estimates of parameters of interest in CRTs can vary widely under different missing data assumptions.

A high proportion of missing data can occur among CRTs because missing data can be found at the individual level as well as the cluster level. In the third paper, I use a simulation study to compare missing data strategies to handle missing cluster level covariates, including the linear mixed effects model, single imputation, single level MI ignoring clustering, MI incorporating clusters as fixed effects, and MI at the cluster level using aggregated data. The results show that when the ICC is small (ICC ≤ 0.1) and the proportion of missing data is low ($\leq 25\%$), the mixed model generates unbiased estimates of regression coefficients and ICC. When the ICC is higher (ICC > 0.1), MI at the cluster level using aggregated data performs well for missing cluster level covariates, though caution should be taken if the percentage of missing data is high.

The outline for this dissertation is as follows. Section 1 presents an in-depth background regarding CRTs, missing data, and pattern mixture models. Section 2 describes the first paper entitled "Statistical analysis and handling of missing data in cluster randomized trials". Section 3 describes the second paper, "A pattern-mixture model approach for handling missing outcome data in longitudinal cluster randomized trials". Section 4 describes the third paper entitled, "Comparison of strategies to impute missing cluster level covariates: a simulation study". Section 5 provides an overall summary and plans for future work. The manuscripts for each paper and their corresponding supplementary files are included as appendices for reference.

1 REVIEW OF THE LITERATURE

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 known to be more appropriate for family-based nutrition interventions, school-based smoking prevention interventions, or interventions aimed at improving obstetric care in hospitals. Cluster-level allocation is often adopted to minimize treatment group contamination, enhance participant compliance, or reduce cost compared to individual randomized trials. Allocation at the group level may also be desirable if individual randomization is unsuitable, unethical, or not feasible [2, 3, 4, 5].

1.1.1 Statistical implications for cluster randomized trials

Cluster level allocation generates several issues for design and statistical analysis. Patients cannot be assumed to be independent because of the similarity among patients within the same cluster, which leads to a reduction of statistical power compared to individual randomized trials. The intracluster correlation coefficient (ICC) is the statistical measure of this cluster dependence, and is defined as the proportion of total variance due to between-cluster variation. 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. The coefficient of variation (CV) is an alternate measure of between-cluster variability, and is defined by the ratio of the standard deviation of cluster sizes to the mean cluster size [4].

When calculating an appropriate sample size for CRTs, decisions must be made for both the number of clusters as well as the number of individuals per cluster. The number of clusters is more important since clusters are the unit of analysis. CRTs with a small number of clusters should be avoided because it may lead to issues in the performance of statistical methods, sample size estimation, and balance of cluster characteristics across treatment arms [6]. For the number of individuals per cluster, a rule of thumb given by Donner and Klar [7] is that power does not substantially increase once the number of individuals per cluster is larger than $1/\rho$. For example, if the expected ICC for a CRT was 0.02, the number of individuals per cluster does not need to exceed $1/0.02 = 50$.

Suppose some variable y was measured on n individuals divided into k clusters. Given each cluster contains $m = n/k$ observations, one can compensate for the cluster design in sample size calculations by multiplying the sample size by a variance inflation factor (also known as the design effect), $1 + (m - 1)\text{ICC}$, so that the sample size is increased to have the same statistical power as an individual design [8]. This shows that even a small ICC with a large number of observations per cluster can lead to underestimated standard errors.

Furthermore, the ICC is independent of the number of observations per cluster. For these reasons, there is more power with a larger number of clusters with fewer observations per cluster than having a few clusters with many observations per cluster, regardless of the ICC [9].

Failing to account for clustering in statistical analysis can lead to falsely low p-values, narrow confidence intervals, and an increased risk of obtaining significant results when there is none, leaving researchers to believe the intervention is more effective than it really is [1, 10] Thus, standard statistical methods are no longer appropriate since the analysis of CRTs need to take into account clusters. Two standard approaches to analyze CRTs include analysis at the cluster level and analysis at the individual level.

1.1.2 Analysis at the cluster level

Cluster level analysis involves reducing all observations within a cluster to a single summary measure, such as a cluster mean or proportion. Standard statistical tests (e.g. t-tests, linear regression models) can then be performed since each data point can now be considered independent [1]. If cluster sizes are not equal, weighting by cluster size for analysis may be desirable, but can lead to reduced power. Inverse variance weighting was proposed to overcome such problems and has been shown to give a more precise estimate for the grand mean [11]. Even though cluster level analysis solves the problem of dependent data, reducing observations to single summary statistics leads to a reduction in sample size and as a result, statistical power.

1.1.3 Analysis at the individual level

Modeling techniques incorporating individual-level covariates in cluster level analysis, such as generalized linear mixed models (GLMM) and generalized estimating equations (GEE), have also been developed [12, 13]. Mixed models are one way to accommodate non-independent data by modeling, or at least taking into account the covariance structure of the data. GEE and random effects logistic regression are two commonly used individual-level analysis methods for estimating the population average and cluster specific intervention effects, respectively, for CRTs with binary outcomes. Random-effects models incorporate cluster-specific random effects into the regression model and assume that the random effects follows a normal distribution [14]. GEE is an extension of generalized linear models to a regression setting with correlated observations. GEE treats the covariance structure as a nuisance variable and is not concerned about the variance of the data [12]. A drawback of the GEE approach is that statistical inference is only valid with a large number of clusters.

These approaches explicitly involve intracluster correlations in the modeling process, which creates a more realistic model of the clustered data. An advantage of these types of models is the ability to control for confounding and reduce bias. However, drawbacks of this approach are that they are more computationally intensive and require a higher sample size of relatively large clusters [2, 8, 15].

1.1.4 Reporting CRTs

Many journals require clinical trials to adhere to the Consolidated Standards of Reporting Trials (CONSORT) statement. The CONSORT statement provides guidelines of items to include in the published report for transparency of methods and results. An extension of the CONSORT statement was developed specifically for CRTs in order to improve conduct and reporting of CRTs. Some items to report in CRTs include the reasons for using a cluster design, how the clustering effect was incorporated into the sample size calculations and analysis, as well as a flow diagram containing the number of clusters and individuals included from randomization to analysis [16].

1.2 Missing data

1.2.1 Implications of missing data

Missing data are common in clinical trials and should be accommodated with appropriate statistical techniques, as they lead to a reduction of power, compromise the advantages created by randomization, and are a potential source of bias. Bias comes from an unaccounted association between the indicator of missing values and the outcome, which cannot be addressed by simply increasing the sample size [17]. In practice, there will almost always be some missing data. Recent reviews of missing data in individual randomized clinical trials have found that the majority have some sort of missing outcome data [18, 19, 20, 21, 22].

1.2.2 How much missing data is acceptable?

There are no set guidelines on how much missing data is reasonable for a clinical trial, since the issue of bias in the treatment effect depends on the missingness mechanism and missing data strategy [23]. Studies with less than 5% missing data will have negligible impact on bias of the treatment effect and power. Schulz and Grimes suggest that missing data greater than 20% is a cause for concern about the validity of the study [24]. Fairclough suggests that drawing inference from studies with 30-50% will be limited. Problems of missing data between 10-20% will depend on the subject matter of the trial [25].

1.2.3 Missing data mechanisms

Missing data mechanisms describe the relationships between the probability of missingness and the observed and unobserved data. They are crucial when choosing an appropriate approach to handle missing data, and have been broadly categorized into the following three classes [26]. Data are considered to be missing completely at random (MCAR) if missingness is independent of observed outcomes and covariates. For example, observations may be missing due to equipment failure or because a patient missed a visit due to an unrelated reason to the illness or intervention (e.g. car broke down). However, MCAR is a strong assumption and is not likely in most clinical trials. There is usually an association between the probability of patient withdrawal and the intervention or baseline measurements prior to withdrawal. 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. Lastly, data are missing not at random (MNAR) when the probability of missingness depends on the missing value even after conditioning on the observed data [26].

1.3 Missing data methods in cluster randomized trials

Commonly used methods for handling missing data in CRTs include complete case analysis, single imputation, multiple imputation, and model based approaches.

1.3.1 Complete case

The most common approach for handling missing outcome data is a complete case analysis, which excludes participants with missing data. This method loses precision since information is deleted, and can yield biased estimation if missingness is not independent of the outcome given covariates [27].

1.3.2 Single imputation

Single imputation strategies fill in missing data with a single value. Last observation carried forward (LOCF) is a popular approach in longitudinal studies despite many statisticians strongly advising against the use of this method [17, 28, 29]. LOCF also makes unlikely assumptions about an individual's trajectory and can lead to either under- or over-estimation of treatment effects [28]. Best or worst case methods impute missing data with the best or worst case value, respectively. For CRTs, mean imputation can involve using all data integrated across clusters in each treatment arm or using data from each

cluster (within-cluster imputation) [30]. Regression imputation should account for multilevel data in CRTs. Issues with single imputation include underestimated variance and potentially biased inferences if the underlying assumption is invalid [17]. Under MCAR, Taljaard et al. showed that cluster mean imputation yields valid inferences for CRTs with large cluster sizes and individual level missing data only [30]. Nevertheless, issues with single imputation include underestimated variance and potentially biased inferences if the underlying assumption is invalid [17].

1.3.3 Multiple imputation

Under the MAR assumption, multiple imputation (MI) takes into account uncertainty by replacing each missing value with a set of possible values to create multiple imputed datasets. These datasets are then analyzed using standard statistical procedures for complete data and combined for inference using specified algorithms, which account for both sampling uncertainty as well as model uncertainty. However, single level MI ignores the hierarchical data structure of CRTs. Taljaard et al. [30] investigated MI strategies in the context of CRTs, and found that ignoring clusters in MI only yields acceptable Type I error rates if the ICC is small (ICC < 0.005). However, if ICC is larger, ignoring clusters may result in severe inflation of Type I error.

Multilevel MI incorporates clustering of CRTs into the imputation process via the Gibbs sampler [31]. A further discussion of multilevel MI is detailed in the next Section 6.2.4. Additional approaches to incorporate clustering in MI have been recently considered by some researchers. Andridge [32] investigated MI strategies accounting for clusters with respect to missing continuous outcomes under MCAR, and showed that MI accounting for clusters using random-effects is more appropriate than incorporating clusters using fixed-effects. For CRTs with missing binary outcomes under MCAR, Ma et al. showed that within-cluster and across-cluster MI approaches yield valid results when the percentage of missing data is higher [33, 34]. Within-cluster MI involves carrying out MI separately for each cluster. Across-cluster MI techniques involves random-effects logistic regression, and includes a dummy variable for each cluster in the MI model [33].

1.3.4 Model based methods for MCAR and MAR

There are some statistical approaches to deal with missing data that do not use formal imputation techniques. Likelihood based mixed models are valid for MAR data if the model is specified correctly, while un-weighted GEE are valid under MCAR if there are a large number of clusters [35, 15]. In order to make a valid complete case analysis under the MAR assumption, inverse probability weighting (IPW) weights complete cases with the inverse of their probability of being observed [36]. Patients with a small chance of being

observed are assigned increased weight in order to compensate for those similar patients who are missing [37]. Although IPW is relatively simple to perform with monotone missing data, it is prone to large weights, causing unstable estimates and high variance [17].

1.4 Sensitivity analysis

A sensitivity analysis in CRTs evaluates whether overall conclusions are influenced by different assumptions and variations, such as different analysis approaches, outcome definitions, and outliers. A sensitivity analysis for missing data evaluates the robustness of inferences from alternative assumptions about the missing data mechanism. 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 [38]. A sensitivity analysis for missing data should be performed in trials expecting a larger amount of missing data [17]. A systematic review on missing data in CRTs found the median percentage of individuals with missing outcome data to be 19%, which shows that most CRTs should be performing a sensitivity analysis for missing data [39].

It has been suggested that researchers weaken the missing data assumption from the primary analysis [17]. 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.

1.5 Pattern mixture models

As a result of MNAR, prediction of observations for those who drop out cannot be reliably predicted using observed data prior to dropping out since the distribution differs between observed and missing values [17]. Two main approaches that have been proposed to handle longitudinal MNAR data include selection models [40] and pattern mixture models (PMMs) [41, 42], which differ in the way the joint-distribution of the outcome and missing data process 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, which have lead to PMMs receiving increased attention [43, 44].

PMMs specify the joint distribution through the marginal distribution of the missing data and the conditional distribution of the measurements given missing data. Originally, PMMs

were proposed by Little for repeated measures with dropouts where the MAR assumption is too strong [41]. Individuals are grouped into dropout patterns based on time of drop out. PMMs explicitly model missing data distributions by first identifying different dropout patterns then including parameters in the outcomes model. Those who drop out are assumed to have a different clinical outcome than the observed outcomes of those who remain in the trial. Implicitly, it is assumed that each pattern has its own missing data process. More detail on PMMs is provided in Section 6.

1.5.1 Under-identification

PMMs are under-identified because some parameters cannot be directly estimated from the available data, as they do not provide enough information to derive the distribution of the unobserved responses. Additional assumptions must be made to estimate all parameters in each dropout pattern.

There are several techniques that have been proposed to deal with under-identification [45]. For example, Little proposed identifying restrictions, which link the inestimable parameters to parameters of the observed data model [41, 42, 44]. For monotone dropout, Little (1993) proposed three strategies to identify unknown parameters: complete case missing value (CCMV), neighboring case missing value (NCMV), and available case missing value (ACMV) [41]. CCMV uses data from the completers to impute missing observations in other patterns. In the case that borrowing information from completers is not sensible, it may be beneficial to borrow from other or all possible patterns. NCMV imputes missing observations by using data from the next identified pattern up. As opposed to previous identifying restrictions, which borrow information from only one pattern, ACMV imputes missing observations by using available data from subjects in higher identified patterns.

In a longitudinal trial with several time points, the large number of dropout patterns can be collapsed for simplification. The advantage to this strategy is that the number of parameters decrease[46]. Although this method is simple, there are strong untestable assumptions being made when grouping dropout patterns.

1.5.2 Advantages & disadvantages

PMMs are appealing since it is natural to assume responders and non-responders have different outcome distributions. These models are clear regarding imputation of missing values because within-pattern models specify the predictive distribution directly. Additionally, PMMs are relatively straightforward regarding model checking for the observed data distribution and have clear interpretation of sensitivity parameters.

However, computation and weighted average patterns for models of large numbers of repeated measures can become complex without simplifying assumptions. PMMs can also be computationally difficult for estimating treatment effects because they require an average over missing data patterns. Furthermore, including auxiliary information requires additional modeling [17].

1.5.3 Extensions

Daniels and Hogan provide a detailed example analyzing both continuous and categorical data using PMMs. They fit the models under MAR and expanded to represent departures from MAR [47]. Other extensions involve strategies to deal with both monotone [48] and non-monotone missingness [49]. PMMs have been applied to a wide range of fields including health-related quality of life settings [50] as well as missing data surveys [51].

2 STATISTICAL ANALYSIS AND HANDLING OF MISSING DATA IN CLUSTER RANDOMIZED TRIALS: A SYSTEMATIC REVIEW

2.1 Introduction

Two potential pitfalls with respect to CRTs are handling missing data and not accounting for clustering in the primary analysis. Missing data decreases power and precision, and can lead to bias by compromising randomization. For example, treatment arm imbalance with respect to missing data is likely to introduce bias when the outcome is related to the reason for patient withdrawal. Even if missing outcome data are balanced across treatment arms, differing reasons for the missing outcome can cause bias [17]. Reviews of individually randomized controlled trials have discovered that most trials have some missing outcome data [19, 22]. Few reports have discussed missing data in CRTs, despite its high likelihood and the recognition that it poses a serious threat to research validity, as discussed by the National Research Council and the Patient Centered Outcomes Research Institute [17].

The second difficulty regarding CRTs is accounting for clustering in the primary analysis. Ignoring clustering can lead to confidence intervals that are too narrow and increased type I error rates[10, 1]. In order to account for clustering, analysis can be performed at the cluster level or at the individual level. Cluster level analysis reduces observations within a cluster to an aggregate value, and then analyzes each independent data point [1, 8]. Analyses at the individual level using general linear models (GLMs) account for non-independent observations within clusters through robust standard errors or adjust using the design effect, an inflation factor used to achieve the same power of an individually randomized trial [52]. Modeling techniques such as generalized estimating equations (GEE) [53] and mixed models [13] explicitly involve intracluster correlation in the modeling process, which enables a more realistic model of the clustered data [13, 12]. Although these models can reduce bias by controlling for confounding at the individual level, they require a higher sample size of a large number of clusters [2, 15, 8].

There have been several reviews on methodological aspects of CRTs (see for example, Simpson et al. [54] and Campbell et al. [16], and references therein). Diaz-Ordaz et al. [55] reviewed imputation methods used to handle missing data in CRTs, but did not distinguish whether a complete case analysis, GEE or mixed model was used to handle missing data in the primary analysis, as these approaches provide valid estimates under differing missing data assumptions. Thus, our objective was to provide a comprehensive review of how missing data are being dealt with in CRTs. The primary aims of our review were to:

1. Identify the proportion of CRTs with missing data at the cluster and individual level
2. Examine the analytical approaches for the primary analysis to find out:
   (a) whether missing data was accommodated
   (b) whether clustering was accounted for
3. Identify the proportion of CRTs reporting a sensitivity analysis for missing data

Secondary aims included assessing techniques for achieving balance in CRTs (stratification, matching, or minimization), differences between observed and expected attrition rates and intracluster correlation.

2.2 Methods

This study was a systematic review of a sample of CRTs published between August 2013 and July 2014. Our methodological strategy was based on guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (See Additional file 1 for compliance details) [56]. We have reported a detailed protocol for this study elsewhere [57].

2.2.1 Eligibility criteria

Eligible studies were restricted to CRTs published in English between August 2013 and July 2014. We included all types of CRTs with human participants, including stepped wedge trials that were reported in the databases listed below [58, 59]. We excluded trial protocols, non- or quasi-experimental designs, secondary trial reports, cost-effectiveness reports and studies where no individual level data were collected. We also excluded trials where the primary outcome was survival, as time-to-event analyses handle censored data differently than other types of data.

2.2.2 Literature search and study selection

Two authors electronically searched for studies found in PubMed, Web of Science (all databases), and PsycINFO. Titles and abstracts were searched containing the terms cluster randomized [randomised], cluster and trial, community trial, community randomized [randomised], or group randomized [randomised]. Two independent reviewers screened titles and abstracts, removed duplicates and screened full texts. The reviewers extracted data from each trial using a standardized, pilot tested form.

2.2.3 Analysis

We defined the number of clusters (and participants) in each trial as the number of clusters (and participants) at randomization. We computed the average number of participants per cluster by dividing the number of participants by the number of clusters. We evaluated the degree of missing data and method(s) for handling missing data in the primary analysis for each trial. When multiple primary outcomes were reported, we used the first outcome listed in the methods section. For primary outcomes measured repeatedly, we used the final follow-up time point to calculate the missing proportion, unless a different time point was specified for the primary analysis.

The proportion of clusters with a missing outcome was calculated as the number of entire clusters with a missing outcome (generally due to the entire cluster dropping out) divided by the number of clusters randomized. A similar calculation was carried out for the proportion of participants with a missing outcome. Of those who reported some missing data, we identified the statistical methods used to handle missing data, classified into the following categories: complete case, single imputation (such as worst case or LOCF), MI (single level or multilevel), GEE, mixed model or IPW.

We computed the number of trials that reported performing a sensitivity analysis and determined the method(s) used to deal with missing data in any sensitivity analysis. We quantified the number of trials that weakened the missingness assumption of their primary analysis to perform their sensitivity analysis as suggested by the Panel on Handling Missing Data in Clinical Trials [17].

For each trial, we calculated the proportion of CRTs performing an individual level or cluster level analysis and whether the analysis accounted for clustering. Individual level analyses were categorized into the following groups: basic inferential test (such as t-test or chi-square)/GLM (such as linear or logistic regression), GEE or mixed model. We recorded whether trials accounted for clustering in sample size calculations, and compared observed and expected ICCs (or CVs) with the mean absolute difference. If a range was reported for the ICC (or CV), we used the upper bound.

2.3 Results

Table 1 presents the general characteristics of the included trials. In total, the median number of clusters randomized was 24, with a range of 2 1,552. The median number of individuals included was 688, with a range of 49 117,100. The average number of individuals per cluster ranged from 1 1,105. Of the 65 trials that collected the outcome repeatedly, 36 (55%) used all of the information in the primary analysis by treating the outcome as a repeated measurement, while 29 (45%) were analyzed at a single time point. Forty-four

(51%) trials used balance techniques to ensure balance after randomization.

Table 1: General characteristics of the 86 randomly selected cluster randomized trials published August 2013 - July 2014

	N (%)
Stepped wedge	4 (5)
Pilot/feasibility	4 (5)
Type of outcome	
Quantitative	41 (48)
Binary	37 (43)
Count	8 (9)
How often outcome was collected	
Single	21 (24)
Repeated	65 (76)
How outcome was treated in the primary analysis	
Single	50 (58)
Repeated	36 (42)
Balance methods used in randomization	
Stratification	27 (31)[1]
Matching	14 (16)
Minimization	3 (3)
None	42 (49)
Presented sample size calculation	60 (70)

[1] One trial also used matching and another trial also used minimization.

2.3.1 Description and handling of missing data

Twenty-seven (31%) trials reported having whole clusters missing in the primary analysis (Table 2). Of these, the median amount of clusters missing was 7%, with a range of 0.8 - 51%. See Figure 2 in Appendix B for the histogram displaying the proportions of included individuals with missing outcomes. Eighty (93%) trials reported having some missing data at the individual level. Of these trials, the median amount of missing individual level data was 19%, with a range of 0.5 - 90%.

Table 2: Proportion of clusters with missing outcome at the primary analysis among the 86 trials included in the review

	N (%)
None	59 (69)
$< 10\%$	14 (16)
$> 10\%$	10 (12)
Unclear	3 (3)

The most common approach for handling missing data in the primary analysis was a complete case analysis (44, 55%) (Table 3). Eighteen (22%) trials used mixed models, 6 (8%) carried out single imputation methods, 4 (5%) trials used un-weighted GEE, 2 (2%) trials performed MI, although neither used multilevel methods.

Sixty (70%) trials presented a sample size calculation, of which 28 (47%) accounted for missing data via sample size inflation. Twenty-six of these trials accounted for missing data at the individual level. Two trials also accounted for missing data at the cluster level by including extra clusters in each trial arm. Two trials mentioned sample size inflation, but were unclear if they accounted for missing data at the cluster or individual level.

2.3.2 Sensitivity analysis for missing data

Fourteen (16%) trials reported a sensitivity analysis for missing data (Table 4). Of these, five (36%) used MI (none of which used multilevel strategies), four (29%) used single imputation, three (21%) used a complete case analysis, one (7%) used a mixed model, and one (7%) used a mixed model with IPW. Only five trials weakened the missingness assumption of the primary analysis to carry out their sensitivity analysis, by assuming MCAR in the primary analysis and MAR in the sensitivity analysis.

Table 3: Handling of missing data in primary analysis among the 80 trials who reported missing outcome data

Methods	< 10% missing N = 14	> 10% missing N = 58	Unclear N = 8	Total N = 80
Complete case	10	31	3	44 (55)
Single imputation				
Worst-case	1	2	0	3 (4)
LOCF	0	2	0	2 (2)
Baseline observation carried forward	0	1	0	1 (1)
Multiple imputation	0	2	0	2 (2)
GEE (un-weighted)	3	0	1	4 (5)
Mixed model/hierarchical/multilevel	0	17	1	18 (22)
Other[1]	0	0	1	1 (1)
Unclear	0	3	2	5 (6)

Abbreviations: LOCF, last observation carried forward; GEE, generalized estimating equation.

[1] One trial excluded participants who dropped out or had no baseline value; for those who participated at both time points, LOCF was carried out for a missing primary outcome.

Table 4: Methods for handling missing data in sensitivity analysis in 14 trials

Sensitivity Method	Primary Analysis	N	Total N (%)
Complete case	MI	2	3 (21)
	Mixed model	1	
Single imputation	Complete case	1	4 (29)
	Single imputation	1	
	Mixed model	2	
MI	Complete case	3	5 (36)
	Mixed model	1	
	Unclear	1	
Mixed model	Complete case	1	1 (7)
Mixed model with IPW	Complete case	1	1 (7)

Abbreviations: MI, multiple imputation; IPW, inverse probability weighting.

2.3.3 Accounting for clustering in the primary analysis

The overwhelming majority of trials carried out an individual level analysis as the primary analysis (83, 97%). Mixed models were the most popular primary analysis used for CRTs (45, 52%). Forty-three (96%) of these trials accounted for clustering by adding cluster as a random effect. Of the 22 (26%) trials performing an individual level basic inferential test or GLM, seven accounted for clustering via robust standard errors or design effect adjustment. Fourteen (16%) trials used GEE, with all of them accounting for clustering by using an exchangeable correlation structure. Of these, one reported estimating standard errors of parameters using the jack-knife method because the number of clusters was small (Table 5) [60]. Four (5%) trials carried out a basic inferential test or GLM at the cluster level. Overall, 68 (79%) trials accounted for clustering in the primary analysis.

Table 5: Primary analysis in 86 cluster randomized trials

	Accounted for clustering[1]		Total
Primary Analysis	Yes N (%)	No N (%)	 N (%)
Individual level:			
Basic inferential test/GLM	7 (32)	15 (68)	22 (26)
GEE	14 (100)	0 (0)	14 (16)
Mixed model	43 (96)	2 (4)[2]	45 (52)
Other[3]	0 (0)	1 (100)	1 (1)
Cluster level:			
Basic inferential test/GLM	4 (100)	0 (0)	4 (5)

Abbreviations: GLM, generalized linear model; GEE, generalized estimating equation.
[1] Denominator is the total number of trials performing respective primary analysis
[2] One trial was unclear
[3] Trial used a descriptive analysis as primary analysis

2.4 Discussion

We performed a systematic review to assess how missing outcome data are being handled in CRTs. Of the 86 included CRTs, most reported some missing outcome data in the primary analysis. Among those that reported missing data, the median proportion of individuals with a missing outcome at the primary analysis was 19%. Sixteen percent of trials carried out a sensitivity analysis for missing data, with all of them reporting more than 10% missing data. Only a third of these trials weakened the missingness assumption

from the primary analysis.

Observed missing data rates generally exceeded expected rates, which means that researchers are not accounting enough for attrition in sample size calculations or adequately following up on participants. Furthermore, only about half (55%) of the trials with repeated measurements used all of the outcome data in the primary analysis. Reducing repeated data to a single time point often generates a strong MCAR assumption and may reduce power. Even if the primary outcome of interest is at a particular time point, previous literature has shown that utilizing all of the information collected can minimize bias due to missing data [61].

In comparison to Diaz-Ordaz et al.s [55] review, we found a higher proportion of trials reporting missing data at the cluster (28% vs. 18%) and individual levels (93% vs. 48%). This may be due to differences in definitions of missing data or because Diaz-Ordaz was not able to verify the amount of missing data in 31% of trials. We observed a similar median cluster attrition rate (7% vs. 10%) and a slightly higher median individual attrition rate (19% vs. 13%). Of the 95 trials with missing data, Diaz-Ordaz et al. found 66% of trials reporting a complete case analysis , GEE or likelihood-based hierarchical/mixed model, while 18% used single imputation, and 6% used MI. Lastly, we found a slightly higher proportion of trials reporting a sensitivity analysis for missing data (16% vs. 11%). Compared to Bell et al.s [22] review of 77 individually randomized controlled trials from 2013, we found a similar proportion of trials reporting missing data (93% vs. 95%). However, CRTs were subject to higher individual level missing data rates (median 19%, up to 90%) compared to individually randomized trials (median 9%, up to 70%). Compared to the individually randomized trials, we found a higher proportion using complete case analysis (55% vs. 45%) and mixed models (22% vs. 15%). Furthermore, we found a similar proportion using GEE (4% vs. 5%) and a lower proportion using single imputation (8% vs. 27%) and MI (2% vs. 8%)

More sophisticated methods are being used. Compared to a review conducted by Simpson et al. [54] of 21 CRTs from 1990-1993, the proportion of trials that took clustering into account in the primary analysis increased over time (57% to 78%). Compared to Scott et al.s [62] review of 150 individually randomized trials in 2001 we found a higher percentage of CRTs using stratification (31% vs. 13%) and a similar percentage using minimization (3% vs. 4%) compared to individually randomized trials.

Our study has several strengths. Eligible studies were all CRT designs including stepped wedge and feasibility studies. In order to minimize the potential for bias during the review process, we had pre-specified search, study selection, and data collection strategies, all of which were carried out by two independent reviewers. We did not limit our sample space to journals with a high impact factor, thereby increasing generalizability. Three independent reviewers performed pilot testing on several trials to create a standardized data collection template. Our study has limitations as well. For example, we only chose CRTs

published in English, which may result in selection bias. It was difficult to identify all CRTs because many do not include cluster as a term in the title or abstract. However, our search strategy included other frequently used terms for cluster randomization such as community randomized and group randomized. Furthermore, we may have underestimated the amount of missing data because we used the CONSORT flow diagram, which may primarily report outcome sample size only. It is possible that missing covariates in regression models resulted in additional missing data and actual smaller sample sizes. Although some trials adjusted for additional covariates beyond balance variables, nearly all were baseline covariates such as age and gender.

In conclusion, missing data are present in the majority of CRTs, yet handling missing data in practice remains suboptimal. Appropriate methods to handle missing clustered data, particularly under the MAR assumption, should be made more accessible by methodological statisticians. Moreover, researchers and applied statisticians should keep up-to-date with such methods in order to increase statistical power in trials and reduce the potential for bias. Thus, we present the following recommendations for CRTs: (1) Attempt to follow up on all randomized clusters and individuals in order to limit the extent of missing data; (2) Perform a primary analysis that is valid under a plausible missingness assumption and that uses all observed data; (3) Perform sensitivity analyses that weaken the missing data assumption to explore the impact of departures made in the primary analysis; (4) Follow the CONSORT extension for cluster trials statement to ensure comprehensive reporting and transparency of methods [17, 38].

3 A PATTERN-MIXTURE MODEL APPROACH FOR HANDLING MISSING OUTCOME DATA IN LONGITUDINAL CLUSTER RANDOMIZED TRIALS

3.1 Introduction

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 [17]. For this reason, modeling dropout might be necessary in order to obtain correct inferences [43]. 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 at the individual level, in which individuals are observed until they drop out and their data from that time point until the end of the study is unobserved.

A sensitivity analysis for missing data is important in CRTs, as it evaluates the robustness of results based on differing missing data assumptions. Despite recommendations to weaken the missing data assumption from the primary analysis [17], a recent review evaluating handling of missing data in CRTs [39] 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 [30, 32, 33, 34], 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.

When analyzing MNAR data, the joint-distribution of the outcome and missing data process is of interest. Pattern mixture models (PMMs) factorize 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 are more easily understandable to applied researchers and clinicians working on clinical trials because the observed data distribution and prediction distribution of missing data are explicitly separated [17, 40].

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 [43, 44]. One 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. 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 [31]. 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.

3.2 Methods

3.2.1 Linear mixed effects model

Consider a CRT with $i = 1, ..., N$ clusters, $j = 1, ..., n_i$ individuals per cluster, and $k = 1, ..., t_{ij}$ measurements per individual. Let $Time_{ijk}$ denote the time of the kth 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}:

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

where $\gamma_i \sim N(0, \sigma^2_\gamma)$ is the random effect at the cluster level and represents deviation of each cluster from the grand mean, $\nu_{ij} \sim N(0, \mathbf{\Sigma}_\nu)$ is the random effect at the individual level and represents deviation of each individual from the cluster effect, and $\epsilon_{ijk} \sim N(0, \sigma^2_\epsilon)$ are the measurement errors terms. Furthermore, γ_i, ν_{ij}, and ϵ_{ijk} are assumed to be uncorrelated.

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

$$\boldsymbol{Y_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\nu_i} + \boldsymbol{\epsilon_i}, \tag{2}$$

where $\boldsymbol{Y_i}$ 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 $N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I_{N_i}})$, where $\boldsymbol{I_{N_i}}$ represents the $N_i \times N_i$ identity matrix.

3.2.2 Pattern mixture models

Little proposed PMMs for repeated measures with dropouts where the MAR assumption is too strong. Let $\mathbf{R}$ be the vector of missingness indicators for the response vector $\mathbf{Y}$ with $\mathbf{Y_{obs}}$ and $\mathbf{Y_{mis}}$ 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 process by:

$$p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \mathbf{R}|\mathbf{X}) = p(\mathbf{R}|\mathbf{X})p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{R}, \mathbf{X}), \tag{3}$$

where $p(\mathbf{R}|\mathbf{X})$ is the conditional probability distribution of the dropout pattern given observed covariates and $p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{R}, \mathbf{X})$ is the probability distribution of the response vector given the dropout pattern and observed covariates [41].

3.2.3 Transforming MAR imputed values to create MNAR imputed values

We employ multilevel MI (described below) and multiply MAR imputed values by a sensitivity parameter k to generate MNAR imputed values such that [63]

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

This creates MNAR observations because the missing data of the non-responders are systematically higher or lower than the observed data of the responders.

3.2.4 Multilevel multiple imputation

Multilevel MI applies the Gibbs sampler to impute missing data found in hierarchical data. 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 $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\nu}, \sigma^2)$
2. Sample $\boldsymbol{\nu}$ from $p(\boldsymbol{\nu}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \sigma^2)$
3. Sample $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma}|\boldsymbol{\nu})$
4. Sample σ^2 from $p(\sigma^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\nu})$
5. Repeat steps 1-4 until convergence

6. Sample $\boldsymbol{y_{mis}}$ from $p(\boldsymbol{y_{mis}}|\boldsymbol{y_{obs}}, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}, \sigma^2)$

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

$$\boldsymbol{\epsilon_i^*} \sim N(0, \sigma^2)$$
$$\boldsymbol{y_i^*} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\nu_i} + \boldsymbol{\epsilon_i^*}$$

where the parameters on the right side of the equation are replaced by values drawn under the Gibbs sampler described above [31].

3.2.5 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 [63]. Let $\hat{Q}_l$ and $\hat{W}_l$ be the point and variance estimates, respectively, obtained from $l = 1, ..., m$ 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 degrees of freedom $v = (m-1)(1 + r^{-1})^2$. The degrees of freedom depends on m and the ratio $r = (1 + m^{-1})B/\bar{W}$, which is the relative increase in variance due to missing data [63].

3.3 Simulation study

3.3.1 Data generation

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

[64]:

$$\begin{aligned} y_{ijk} =& \beta_0 + \beta_1(Time_{ijk}) + \beta_2(Trt_i) + \beta_3(Trt_i \times Time_{ijk}) \\ &+ \beta_4(Drop_{ijk} \times Trt_i \times Time_{ijk}) + \gamma_i + \nu_{ij} + \epsilon_{ijk} \end{aligned} \tag{5}$$

where $\gamma_i \sim N(0, \sigma^2_\gamma)$ denotes the random cluster effect, $\nu_{ij} \sim N(0, \mathbf{\Sigma}_\nu)$ denotes the random individual effect, and $\epsilon_{ijk} \sim N(0, \sigma^2_\epsilon)$ 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. The regression coefficients were defined as $\beta_0 = 7$, $\beta_1 = -1$, $\beta_2 = 0$, $\beta_3 = -2$, $\beta_4 = 3$. 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. The total number of clusters and cluster size varied in pairs as (12, 30), (12, 100), (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 a true $k = 1.0$ for the control arm and $k = 1.75$ for the treatment arm, and creates MNAR data in the treatment arm when the data are deleted.

3.3.2 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 [65]. 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 $k = 1.0$ (MAR). For the treatment arm, we multiplied each imputed value by $k = (0.8, 1.0, 1.3, 1.7)$, which decreases the imputed values by 20%, and increases the imputed values by 0%, 30%, and 70%.

Using the completed dataset, we modeled the outcome with a mixed model (`lme4`) using Equation 5, 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 5, the true change over time in the treatment arm was,

$$([0.60 \times (\beta_0 + \beta_1 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_3)]) - \beta_0 = -1.8$$

and the true treatment effect was,

$$([0.60 \times (\beta_0 + \beta_1 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_3)]) - (\beta_0 + \beta_1) = -0.8.$$

Parameter estimates were pooled using Rubin's rules as implemented in `mice` [63]. 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. Empirical standard error: standard deviation of mean across samples
4. Ratio of model-based to empirical standard error

3.3.3 Results

We present the results of our simulations in Tables 6 - 9. Table 6 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. Under the MAR assumption ($k = 1.0$) and the incorrect MNAR assumption ($k = 0.8$), the estimates have a severe downward bias. Percent bias is more extreme in the treatment effect.

Table 6: 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	30	0.001	-82.77	-64.94	-38.20	-2.55	-186.79	-146.68	-86.52	-6.31
		0.01	-83.21	-65.39	-38.66	-3.02	-189.77	-149.67	-89.53	-9.34
		0.1	-85.04	-67.47	-41.13	-6.01	-195.05	-155.54	-96.26	-17.24
		0.3	-84.75	-66.68	-39.59	-3.46	-180.70	-140.06	-79.09	2.21
		0.5	-84.37	-66.75	-40.32	-5.07	-182.25	-142.61	-83.14	-3.84
12	100	0.001	-83.75	-65.91	-39.14	-3.45	-186.73	-146.58	-86.36	-6.06
		0.01	-83.51	-65.55	-38.61	-2.69	-183.83	-143.42	-82.80	-1.98
		0.1	-84.53	-66.73	-40.05	-4.47	-185.90	-145.87	-85.82	-5.76
		0.3	-84.15	-66.48	-39.98	-4.64	-195.72	-155.97	-96.34	-16.83
		0.5	-85.24	-67.26	-40.28	-4.32	-189.63	-149.17	-88.48	-7.56
30	30	0.001	-83.96	-66.08	-39.26	-3.51	-188.81	-148.58	-88.24	-7.78
		0.01	-85.99	-68.29	-41.74	-6.33	-190.47	-150.63	-90.88	-11.22
		0.1	-84.41	-66.66	-40.05	-4.56	-192.69	-152.76	-92.87	-13.02
		0.3	-83.97	-66.55	-40.42	-5.57	-198.89	-159.69	-100.89	-22.49
		0.5	-84.74	-66.89	-40.13	-4.44	-204.15	-163.99	-103.77	-23.46
30	100	0.001	-83.93	-66.20	-39.61	-4.15	-190.23	-150.33	-90.50	-10.71
		0.01	-84.35	-66.61	-40.01	-4.54	-190.29	-150.38	-90.53	-10.72
		0.1	-84.87	-66.98	-40.14	-4.36	-184.42	-144.16	-83.78	-3.27
		0.3	-84.28	-66.29	-39.31	-3.34	-183.15	-142.69	-81.99	-1.05
		0.5	-83.57	-65.76	-39.03	-3.39	-190.36	-150.27	-90.14	-9.95

Abbreviations: MNAR, missing not at random; ICC, intracluster correlation coefficient
[1]Treatment effect: mean difference between treatment arms at follow-up

Table 7 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 $k = 1.7$. Furthermore, coverage for both estimates decreases as ICC increases as seen under $k = 1.7$.

Table 7: 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	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	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, intracluster correlation coefficient

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

Tables 8 - 9 display the empirical standard errors and ratios of model-based to empirical standard errors for change over time in the treatment arm and treatment effect. Overall, results were similar for the percent bias of the treatment arm change over time and treatment effect. Larger k overestimates the standard errors because the imputed values are multiplied, which increases variances of the estimates.

Table 8: Empirical standard errors for change over time in the treatment arm and treatment effect.

No. clusters	Cluster size	ICC	Treatment arm change over time	Treatment effect[1]
12	30	0.001	0.224	0.335
		0.01	0.234	0.371
		0.1	0.234	0.666
		0.3	0.231	1.147
		0.5	0.234	1.848
12	100	0.001	0.128	0.182
		0.01	0.122	0.245
		0.1	0.129	0.609
		0.3	0.125	1.214
		0.5	0.126	1.781
30	30	0.001	0.364	0.508
		0.01	0.372	0.592
		0.1	0.365	1.078
		0.3	0.366	1.984
		0.5	0.386	2.847
30	100	0.001	0.212	0.307
		0.01	0.204	0.416
		0.1	0.203	0.936
		0.3	0.201	1.778
		0.5	0.190	2.757

Abbreviations: ICC, intracluster correlation coefficient

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

Table 9: 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	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	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, intracluster correlation coefficient

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

3.4 Application to the PoNDER study

3.4.1 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 [66]. 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. 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 10 displays the means and standard deviations of EPDS score by treatment arm at baseline.

Table 10: PoNDER study. Means and standard deviations of baseline EPDS score by treatment arm and dropout pattern.

	Control N = 1151		Treatment N = 2268	
Dropout pattern	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

3.4.2 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 in Appendix C 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 treat-

ment arm and (2) treatment effect, defined as the mean difference in arms post-treatment. Inferences were combined using Rubin's rules.

3.4.3 Results

Table 11 displays the results of each PMM scenario. As k increases the slope of the treatment arm as well as the treatment effect attenuate. 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.

Table 11: 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
[1]Treatment effect: mean difference between treatment arms at follow-up

3.5 Discussion

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.

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 [67]. This shifts the imputed values of the non-responders, while preserving the standard errors of the estimates of interest. One way to shift imputed values is to identify the δ that increases the average MAR imputed values by a certain percentage. In our PoNDER example, the average MAR imputed EPDS score among the non-responders in the treatment arm was 5.73. If we wanted to increase the imputed values by an average of 30%, we compute $\delta = 5.73 \times 0.3 = 1.72$. Rather than multiplying imputed values by $k = 1.3$, we add $\delta = 1.72$ to all imputed values. By adding δ, the imputed values are shifted without inflating the standard errors. 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 [68]. A range of plausible k or δ can be specified, or an average can be specified if a single analysis is preferred.

Multiplying imputed MAR values by k to create MNAR values can be implemented in both treatment arms or in one treatment arm only. 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. Extending the PMM to a CRT with more than two time points becomes more challenging. The multiplier k can be specified at each time point 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.

One limitation is that our PMM analysis was performed in concordance with our simulations. We did not see how parameters of interest and their corresponding standard errors would be affected if we misspecified the model. It would be beneficial to evaluate robustness to model misspecification since the true distribution of the data is not usually known in practice.

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.

4 COMPARISON OF STRATEGIES TO IMPUTE MISSING CLUSTER LEVEL COVARIATES: A SIMULATION STUDY

4.1 Introduction

Empirically, the ICC is very low for clustered data where subjects are grouped within clusters. For example, Bell and McKenzie [27] assessed 87 ICCs from 15 psycho-oncology cluster randomized trials, which randomize clusters to treatment arms rather than individuals, and found the median ICC to be 0.02. Adams et al. [69] examined 1039 ICCs from 31 multilevel studies found in primary care research, and reported the median ICC to be 0.01. Along with subjects grouped within clusters, multilevel data also occur in longitudinal data, where individuals are measured repeatedly over time. ICCs found within longitudinal data are generally higher. We focus our attention on non-longitudinal multilevel data where subjects are grouped within clusters and the ICC is lower.

The proportion of missing data can be higher in multilevel data compared to independent data because missing data can occur at the individual level and cluster level. In an educational dataset where students are grouped by classroom, missing data can occur at the student (or individual) level, such as age of the student or test score. Missing data can also occur at the classroom (or cluster) level, such as teacher's highest level of education. Missing individual level outcomes and covariates have been considered by many. Some references include [30, 32, 31, 33, 34]. However, missing data among cluster level covariates have received limited attention. Gibson et al. [70] and Cheung [71] both evaluated strategies to handle continuous missing cluster level covariates under MCAR, such as complete case analysis, single level multiple imputation (MI) ignoring clustering, and mean substitution. Both studies found complete case analysis to perform well with less than 50% missing cluster level data. However, a complete case analysis can be very inefficient when missing cluster level covariates are present, since all of the observations within the cluster are removed. Additionally, this method loses precision since information is being deleted, and can produce biased estimates if the missing data mechanism is not MCAR. Both studies found single level MI ignoring clustering to be a poor strategy when implemented within the multilevel structure, because it produces underestimated standard errors.

When imputing cluster level variables, imputed values must be the same within each cluster. Gelman and Hill [72] suggested an approach using MI to impute missing cluster level covariates, which involves separating the data into subject level and cluster level datasets and imputing within each dataset. The imputed data are then combined to create complete datasets and analyzed for inference. This method, however, has not yet been compared to other commonly used techniques to impute missing cluster level data, such as the linear mixed effects model, mean substitution, and other MI approaches. We extend the inves-

tigation of missing cluster level covariates by performing a simulation study to assess the performance of missing data strategies under MAR, as this assumption may be more reasonable in practice. We examine the sensitivity of methods to the total number of clusters, cluster size (number of subjects per cluster), ICC, and percentage of missingness. Since cluster level covariates are often found to be categorical, we evaluate these methods when the missing cluster level covariate is categorical as well as continuous.

4.2 Methods

We used a simulation study designed after Van Buuren's previous work in missing multilevel data [31] to investigate the performance of strategies to handle missing cluster level covariates under the MAR assumption. This section contains an overview of commonly used methods to accommodate missing continuous and categorical cluster level data, followed by an introduction to the linear mixed effects model, which we use to compute regression coefficients and variance components after performing each missing data method.

4.2.1 Missing data methods

There are several methods that can be used to handle missing cluster level covariates, including the linear mixed effects model, mean substitution (continuous variable), mode substitution (categorical variable), single level MI ignoring clustering, fixed effects MI, and MI aggregate imputation.

The linear mixed effects model (mixed model) uses a likelihood-based approach to estimate parameters. Under the correct model specification, the mixed model performs an implicit imputation and produces unbiased estimates under MAR if outcome data are missing. This method excludes observations with missing cluster level covariates from the analysis, and can produce biased estimates in the presence of missing covariate variables even under MCAR [73].

Among continuous cluster level covariates, mean substitution replaces missing observations with the overall mean across observed clusters. For categorical variables, mode substitution replaces missing data with the most common category found among the observed data. Similar to other single imputation techniques, which replace missing observations with a single value, these methods are prone to underestimation of variance. Although simple, these methods do not condition on any other information in the observation, which can generate misleading relationships between variables [31].

The MI procedure can be separated into two steps: (1) imputation of missing data and (2) analysis of complete multiply imputed datasets. We carry out MI using the `mice` package in R, which uses the multivariate imputation by chained equations (MICE) technique [74].

Imputation models are used for each variable with missing data so that unobserved values are imputed based on a conditional distribution of other variables with observed data to create m complete datasets. Each completed dataset is then analyzed using standard statistical techniques and combined for inference [63]. Standard MI, which we will call single level MI, assumes the observations are independent and ignores the clustered structure found in multilevel data. This results in underestimated standard errors and confidence intervals that are too narrow [31].

One approach to incorporate clustering in the MI procedure, that we will call fixed effects MI, includes cluster as a fixed effect in the imputation model when performing MI, which models the differences in intercepts between the clusters [31]. However, single level MI and fixed effects MI will not impute the same value for all individuals within a single cluster, which is inappropriate when imputing missing cluster level variables. Gelman and Hill [72] proposed using a MI aggregate imputation approach, which imputes the same value for each individual within a cluster. This method involves separating the data into two datasets: one individual level dataset and one cluster level dataset. In order to impute missing cluster level data, the individual level data are first aggregated into cluster level summaries (such as the cluster mean). The aggregated individual level data are then incorporated with the cluster level data for imputation, so that the combined dataset includes a single row for each cluster with the cluster level variables and the aggregated individual level variables. The aggregated individual level variables are included in the imputation model to impute a single value for each missing cluster level covariate. The imputed cluster level covariates are then combined with the original individual level data, so that the final dataset includes the cluster level variables with the same imputed value across individuals within a cluster. Each completed multiply imputed dataset is then analyzed and combined for inference [72].

4.2.2 Linear mixed effects model

The linear mixed effects model is often used to analyze clustered data. We use this method to analyze the completed dataset after performing each imputation approach. Consider a multilevel dataset with $i = 1, ..., N$ clusters and $j = 1, ..., n_i$ subjects per cluster. A single outcome of interest Y_{ij} is modeled by:

$$Y_{ij} = X_{ij}\beta + U_i + e_{ij} \tag{6}$$

where Y_{ij} is an $n_i \times 1$ vector of responses, X_{ij} is a known $n_i \times p$ design matrix of fixed effects at the individual level or cluster level, β is a $p \times 1$ vector of unknown fixed effects, U_i represents the random cluster effects distributed $N(0, \sigma_B^2)$, and e_{ij} represents the individual error terms distributed $N(0, \sigma_W^2)$. Additionally, U_i and e_{ij} are assumed to be uncorrelated. The variance of Y_{ij} is $\sigma^2 = \sigma_B^2 + \sigma_W^2$, where σ_B^2 denotes the between cluster variance and σ_W^2 denotes the within cluster variance. The ICC $= \sigma_B^2/\sigma^2$.

4.3 Simulation study

In this section, we describe our simulation study to assess methods to handle missing continuous and categorical cluster level covariates.

Continuous cluster level covariate

Multilevel data with a continuous cluster level covariate, W_i, were generated using the following model [31]:

$$Y_{ij} = \beta_0 + \beta_1 W_i + U_i + e_{ij} \tag{7}$$

with $U_i \sim N(0, \sigma^2_W)$ and $e_{ij} \sim N(0, \sigma^2_B)$ denoting the random cluster effects and measurement error terms, respectively. The regression coefficients were defined as $\beta_0 = 0$ and $\beta_1 = 0.5$. We set the variance parameters $\sigma^2 = \sigma^2_W + \sigma^2_B = 0.75$, and varied the ICC = (0.001, 0.01, 0.1, 0.3). We simulated both small and large sample sizes by varying the total number of clusters $N = (24, 60)$ and cluster size $n_i = (20, 50)$.

The cluster level covariate W_i was deleted under the MAR assumption with 25% and 50% missing. For the cluster average Y_i less than than the upper 25th (or 50th) percentile of the standard normal distribution, the non-response probability in W_i was 10%. For the cluster average Y_i greater than or equal to the upper 25th (or 50th) percentile of the standard normal distribution, the non-response probability in W_i was 90%.

We simulated 1000 replications from each scenario described above, and performed the following methods to handle missing cluster level continuous covariates: the mixed model (MM), mean substitution (MN), single level MI ignoring clustering (SL), fixed effects MI (FE), MI aggregate imputation (AG). All MI methods were carried out as described in the previous section using the `mice` package in `R` version 3.2.3 [65]. Imputations were fixed to 20 with the outcome variable included in the imputation model. For each completed dataset, we modeled the outcome with a mixed model via the R package `lme4` using Equation 7. Fixed and random parameter estimates were pooled using Rubin's Rules [63] as implemented in `mice`. For each missing data method, we calculated the estimated fixed and random effects parameters and examined their bias, defined as the difference between the estimate of the parameter and the true value. Coverage was also computed as the proportion of times the 95% confidence interval contained the true value of the fixed parameter estimate.

Categorical cluster level covariate

When the cluster level covariate was categorical, we simulated the data similarly, except with the following multilevel model:

$$Y_{ij} = \beta_0 + \beta_1(\text{Group 2}) + \beta_2(\text{Group 3}) + U_i + e_{ij}$$

where the regression coefficients were defined as $\beta_0 = 0$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. With β_1 and β_2 set as dummy variables, this creates three cluster level groups with means 0 (group 1), 0.5 (group 2), and 0.9 (group 3). There was an equal probability of being assigned to one of the three groups. The following methods used to handle missing cluster level categorical data were examined: the mixed model (MM), mode substitution (MD), single level MI ignoring clustering (SL), fixed effects MI (FE), MI aggregate imputation (AG). We deleted the categorical cluster level covariate under MAR and assessed performance of methods similar to the continuous cluster level covariate case.

4.4 Results

Continuous cluster level covariate

The results of our simulations for missing data in the continuous cluster level covariate are presented in Tables 12 - 14. When the ICC was small (ICC ≤ 0.1) with 25% missing data, the mixed model performed best, as it yielded unbiased regression coefficients and reasonable coverage rates. Compared to the other missing data strategies, the mixed model generated the closest ICC estimates to the true value, particularly when the ICC was lower. However, when the ICC was higher (ICC > 0.1), the mixed model produced severely underestimated regression coefficients and performed worse with an increased amount of missing data.

MI aggregate imputation performed best when the ICC was larger. It generated consistently close fixed parameter estimates to the true values as well as adequate coverage of the fixed parameters. Similar to the other imputation approaches, MI aggregate imputation tended to overestimate the ICC particularly with 50% missingness, though it generally performed better.

The worst method was mean substitution, which became more apparent with more missing cluster level covariates. Mean substitution severely overestimated the β_0 coefficient as well as the ICC. Single level MI and fixed effects MI performed better than mean substitution, but were under-coveraged for the fixed effects due to underestimation of the standard errors. The results did not change substantially when varying the total number of clusters or cluster size.

Table 12: Estimates of the regression coefficients based on methods to handle 25% and 50% missing data in continuous cluster level covariates

No. clusters	Cluster size	ICC	$\beta_0 = 0$ MM	MN	SL	FE	AG	$\beta_1 = 0.5$ MM	MN	SL	FE	AG
25% missing												
24	20	0.001	-0.01	0.08	0.07	0.07	0.05	0.48	0.48	0.48	0.48	0.53
24	20	0.01	-0.02	0.08	0.07	0.07	0.05	0.47	0.47	0.47	0.47	0.53
24	20	0.1	-0.05	0.08	0.06	0.06	0.05	0.43	0.43	0.39	0.39	0.49
24	20	0.3	-0.14	0.07	0.03	0.03	0.04	0.38	0.38	0.27	0.27	0.43
24	50	0.001	-0.01	0.08	0.06	0.06	0.04	0.49	0.49	0.47	0.47	0.54
24	50	0.01	-0.01	0.08	0.06	0.06	0.04	0.49	0.49	0.46	0.46	0.54
24	50	0.1	-0.05	0.08	0.05	0.05	0.04	0.44	0.44	0.38	0.38	0.5
24	50	0.3	-0.14	0.07	0.02	0.02	0.03	0.40	0.40	0.25	0.25	0.46
60	20	0.001	-0.02	0.08	0.06	0.06	0.03	0.48	0.48	0.48	0.48	0.53
60	20	0.01	-0.02	0.08	0.06	0.06	0.03	0.47	0.47	0.47	0.47	0.53
60	20	0.1	-0.06	0.08	0.05	0.05	0.03	0.43	0.43	0.40	0.40	0.50
60	20	0.3	-0.14	0.08	0.03	0.03	0.03	0.39	0.39	0.26	0.26	0.47
60	50	0.001	-0.01	0.08	0.06	0.06	0.03	0.49	0.49	0.47	0.47	0.53
60	50	0.01	-0.01	0.08	0.06	0.06	0.03	0.49	0.49	0.46	0.46	0.53
60	50	0.1	-0.05	0.08	0.05	0.05	0.03	0.44	0.44	0.37	0.37	0.51
60	50	0.3	-0.14	0.08	0.03	0.03	0.03	0.39	0.39	0.24	0.24	0.47
50% missing												
24	20	0.001	-0.08	0.26	0.20	0.20	0.20	0.45	0.45	0.41	0.41	0.51
24	20	0.01	-0.09	0.26	0.20	0.20	0.19	0.44	0.44	0.41	0.41	0.49
24	20	0.1	-0.19	0.20	0.13	0.13	0.13	0.39	0.39	0.33	0.33	0.41
24	20	0.3	-0.34	0.15	0.05	0.05	0.08	0.35	0.35	0.24	0.24	0.36
24	50	0.001	-0.04	0.29	0.22	0.22	0.23	0.47	0.47	0.42	0.42	0.54
24	50	0.01	-0.06	0.28	0.22	0.22	0.22	0.46	0.46	0.42	0.42	0.53
24	50	0.1	-0.18	0.20	0.13	0.13	0.14	0.39	0.39	0.32	0.32	0.42
24	50	0.3	-0.33	0.15	0.06	0.06	0.08	0.35	0.35	0.23	0.23	0.35
60	20	0.001	-0.07	0.27	0.20	0.20	0.13	0.45	0.45	0.42	0.42	0.53
60	20	0.01	-0.09	0.26	0.20	0.20	0.13	0.45	0.45	0.41	0.41	0.52
60	20	0.1	-0.19	0.21	0.13	0.13	0.10	0.40	0.40	0.35	0.34	0.47
60	20	0.3	-0.33	0.15	0.06	0.06	0.06	0.36	0.36	0.24	0.24	0.42
60	50	0.001	-0.03	0.29	0.22	0.22	0.14	0.48	0.48	0.43	0.43	0.56
60	50	0.01	-0.05	0.28	0.21	0.21	0.14	0.47	0.47	0.41	0.41	0.54
60	50	0.1	-0.17	0.21	0.13	0.13	0.10	0.41	0.41	0.33	0.33	0.48
60	50	0.3	-0.33	0.15	0.06	0.06	0.06	0.36	0.36	0.23	0.23	0.42

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 13: Estimates of the intracluster correlation coefficient and residual variance based on methods to handle 25% and 50% missing data in continuous cluster level covariate

No. clusters	Cluster size	ICC	MM	MN	SL	FE	AG	σ^2	MM	MN	SL	FE	AG
25% missing													
24	20	0.001	0.00	0.11	0.08	0.08	0.04	0.749	0.74	0.75	0.73	0.73	0.74
24	20	0.01	0.01	0.12	0.09	0.09	0.05	0.74	0.74	0.74	0.72	0.72	0.74
24	20	0.1	0.08	0.24	0.19	0.19	0.16	0.65	0.65	0.65	0.64	0.64	0.65
24	20	0.3	0.23	0.47	0.41	0.41	0.39	0.45	0.45	0.45	0.44	0.44	0.45
24	50	0.001	0.00	0.11	0.08	0.08	0.03	0.749	0.75	0.75	0.73	0.73	0.75
24	50	0.01	0.01	0.11	0.08	0.08	0.04	0.74	0.74	0.74	0.72	0.72	0.74
24	50	0.1	0.09	0.23	0.19	0.19	0.15	0.65	0.65	0.65	0.64	0.64	0.65
24	50	0.3	0.24	0.47	0.43	0.43	0.39	0.45	0.45	0.45	0.44	0.44	0.45
60	20	0.001	0.00	0.11	0.07	0.07	0.02	0.749	0.75	0.75	0.73	0.73	0.75
60	20	0.01	0.01	0.12	0.08	0.08	0.03	0.74	0.74	0.74	0.72	0.72	0.74
60	20	0.1	0.08	0.23	0.19	0.19	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	20	0.3	0.23	0.46	0.41	0.41	0.36	0.45	0.45	0.45	0.44	0.44	0.45
60	50	0.001	0.00	0.10	0.07	0.07	0.02	0.749	0.75	0.75	0.73	0.73	0.75
60	50	0.01	0.01	0.11	0.08	0.08	0.03	0.74	0.74	0.74	0.72	0.72	0.74
60	50	0.1	0.09	0.23	0.19	0.19	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	50	0.3	0.23	0.46	0.42	0.42	0.36	0.45	0.45	0.45	0.44	0.44	0.45
50% missing													
24	20	0.001	0.01	0.19	0.16	0.16	0.12	0.749	0.74	0.75	0.71	0.71	0.75
24	20	0.01	0.01	0.21	0.17	0.17	0.14	0.74	0.73	0.74	0.70	0.70	0.74
24	20	0.1	0.07	0.30	0.25	0.25	0.24	0.65	0.65	0.65	0.62	0.62	0.65
24	20	0.3	0.20	0.52	0.45	0.45	0.45	0.45	0.45	0.45	0.43	0.43	0.45
24	50	0.001	0.00	0.19	0.15	0.15	0.12	0.749	0.75	0.75	0.71	0.71	0.75
24	50	0.01	0.01	0.20	0.17	0.17	0.14	0.74	0.74	0.74	0.70	0.70	0.74
24	50	0.1	0.08	0.30	0.25	0.25	0.24	0.65	0.65	0.65	0.62	0.62	0.65
24	50	0.3	0.21	0.51	0.45	0.45	0.46	0.45	0.45	0.45	0.43	0.43	0.45
60	20	0.001	0.00	0.19	0.14	0.14	0.08	0.749	0.74	0.75	0.70	0.70	0.75
60	20	0.01	0.01	0.20	0.15	0.15	0.09	0.74	0.74	0.74	0.69	0.69	0.74
60	20	0.1	0.07	0.29	0.24	0.24	0.19	0.65	0.65	0.65	0.61	0.61	0.65
60	20	0.3	0.21	0.50	0.43	0.43	0.40	0.45	0.45	0.45	0.43	0.43	0.45
60	50	0.001	0.00	0.18	0.14	0.14	0.07	0.749	0.75	0.75	0.70	0.70	0.75
60	50	0.01	0.01	0.19	0.15	0.15	0.08	0.74	0.74	0.74	0.69	0.69	0.74
60	50	0.1	0.08	0.29	0.24	0.24	0.19	0.65	0.65	0.65	0.61	0.61	0.65
60	50	0.3	0.21	0.50	0.43	0.43	0.40	0.45	0.45	0.45	0.43	0.43	0.45

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 14: Coverage of true values by the 95% confidence interval of regression coefficients based on methods to handle 25% and 50% missing data in continuous cluster level covariate

No. clusters	Cluster size	ICC	β_0 MM	MN	SL	FE	AG	β_1 MM	MN	SL	FE	AG
25% missing												
24	20	0.001	95	88	87	87	89	94	100	98	98	96
24	20	0.01	94	87	87	87	90	91	99	96	96	95
24	20	0.1	88	92	92	92	94	86	96	82	80	94
24	20	0.3	79	93	93	93	93	82	93	47	47	94
24	50	0.001	94	87	87	87	89	95	100	97	97	94
24	50	0.01	93	89	90	90	90	92	100	95	96	95
24	50	0.1	89	91	92	92	93	87	97	63	63	97
24	50	0.3	77	93	92	92	95	85	96	20	20	95
60	20	0.001	94	67	72	72	87	90	100	98	98	90
60	20	0.01	90	69	74	74	87	88	100	96	96	92
60	20	0.1	81	81	88	88	90	76	94	61	61	94
60	20	0.3	56	86	92	92	93	73	90	7	8	94
60	50	0.001	94	60	64	64	79	93	100	95	95	79
60	50	0.01	92	66	72	72	83	91	100	89	90	82
60	50	0.1	80	78	87	87	90	81	96	21	22	94
60	50	0.3	55	89	94	94	92	71	90	1	1	94
50% missing												
24	20	0.001	84	68	59	59	64	92	100	89	91	97
24	20	0.01	79	69	59	59	69	89	100	87	86	97
24	20	0.1	54	82	80	80	88	75	98	67	66	94
24	20	0.3	39	92	91	91	94	74	96	42	43	93
24	50	0.001	90	55	36	35	47	93	100	80	79	97
24	50	0.01	81	60	44	44	55	89	100	75	75	97
24	50	0.1	55	82	76	76	88	78	99	47	48	96
24	50	0.3	39	90	89	89	92	76	95	29	30	92
60	20	0.001	64	9	9	8	41	83	100	77	77	91
60	20	0.01	54	13	14	14	48	78	100	70	71	93
60	20	0.1	28	46	56	57	75	66	95	41	40	92
60	20	0.3	12	77	88	89	91	63	92	13	13	89
60	50	0.001	83	2	1	2	28	91	100	60	61	83
60	50	0.01	72	4	3	3	35	82	100	46	46	90
60	50	0.1	30	42	51	50	70	67	97	15	15	92
60	50	0.3	12	76	88	88	92	63	89	7	7	87

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Categorical cluster level covariate

The results of our simulations for missing data in categorical cluster level covariates are presented in Tables 15 - 17. With 25% missing data and lower ICC (ICC ≤ 0.1), the best methods were the mixed model and MI aggregate imputation. Both strategies generated reasonable fixed parameter estimates, coverage, and ICC estimates. The mixed model had problems with convergence when the total number of clusters was lower ($N = 24$). When ICC was higher, MI aggregate imputation outperformed all other strategies. However, when the amount of missing data increased to 50%, none of the missing data strategies performed particularly well. Overall, the worst method to handle missing categorical cluster level covariates was mode substitution. The fixed parameter estimates were extremely biased, especially with higher ICC. Mode substitution also overestimated the ICC, and became worse with 50% missing data.

Table 15: Estimates of the regression coefficients based on methods to handle 25% and 50% missing data in categorical cluster level covariates

No. clusters	Cluster size	ICC	$\beta_0 = 0$					$\beta_1 = 0.5$					$\beta_2 = 0.9$				
			MM	MD	SL	FE	AG	MM	MD	SL	FE	AG	MM	MD	SL	FE	AG
25% missing																	
24	20	0.001	0.00	0.06	0.02	0.03	0.00	0.50	0.45	0.50	0.49	0.51	0.87	0.79	0.87	0.86	0.89
24	20	0.01	0.00	0.06	0.02	0.03	0.00	0.50	0.45	0.50	0.50	0.51	0.86	0.78	0.85	0.85	0.89
24	20	0.1	-0.01	0.11	0.09	0.09	0.01	0.48	0.41	0.46	0.45	0.52	0.77	0.66	0.73	0.74	0.84
24	20	0.3	-0.03	0.17	0.21	0.20	0.07	0.41	0.31	0.31	0.30	0.46	0.66	0.49	0.48	0.50	0.73
24	50	0.001	0.00	0.04	0.01	0.02	0.00	0.50	0.46	0.49	0.49	0.50	0.90	0.83	0.88	0.88	0.90
24	50	0.01	0.00	0.05	0.02	0.03	0.00	0.50	0.46	0.49	0.48	0.50	0.88	0.81	0.86	0.86	0.89
24	50	0.1	0.00	0.11	0.11	0.11	0.02	0.47	0.40	0.43	0.42	0.51	0.78	0.67	0.69	0.69	0.84
24	50	0.3	-0.02	0.16	0.23	0.23	0.06	0.41	0.32	0.28	0.28	0.46	0.67	0.50	0.43	0.46	0.73
60	20	0.001	0.00	0.07	0.02	0.02	0.00	0.50	0.45	0.50	0.50	0.50	0.86	0.78	0.87	0.86	0.90
60	20	0.01	0.00	0.07	0.03	0.03	0.00	0.50	0.45	0.50	0.50	0.51	0.85	0.77	0.85	0.85	0.90
60	20	0.1	0.00	0.14	0.09	0.10	0.01	0.47	0.38	0.45	0.44	0.51	0.76	0.63	0.72	0.72	0.87
60	20	0.3	-0.02	0.22	0.23	0.22	0.04	0.40	0.26	0.29	0.29	0.47	0.66	0.44	0.46	0.48	0.80
60	50	0.001	0.00	0.04	0.01	0.01	0.00	0.50	0.46	0.50	0.49	0.50	0.89	0.82	0.89	0.88	0.90
60	50	0.01	0.00	0.06	0.02	0.02	0.00	0.50	0.45	0.49	0.49	0.50	0.88	0.80	0.87	0.87	0.90
60	50	0.1	0.00	0.13	0.11	0.11	0.01	0.48	0.40	0.43	0.43	0.52	0.78	0.65	0.69	0.69	0.88
60	50	0.3	-0.02	0.21	0.25	0.25	0.04	0.41	0.27	0.27	0.27	0.49	0.67	0.45	0.43	0.44	0.81
50% missing																	
24	20	0.001	0.00	0.38	0.25	0.25	0.15	0.37	0.02	0.34	0.36	0.46	0.86	0.47	0.77	0.76	0.70
24	20	0.01	-0.01	0.39	0.26	0.25	0.16	0.36	0.01	0.33	0.34	0.45	0.83	0.43	0.72	0.72	0.68
24	20	0.1	-0.06	0.36	0.27	0.27	0.19	0.32	-0.02	0.26	0.26	0.36	0.68	0.26	0.53	0.57	0.57
24	20	0.3	-0.19	0.31	0.27	0.27	0.18	0.32	-0.05	0.20	0.20	0.32	0.61	0.13	0.36	0.41	0.50
24	50	0.001	0.00	0.39	0.25	0.24	0.14	0.41	0.04	0.37	0.36	0.50	0.91	0.51	0.81	0.83	0.72
24	50	0.01	0.00	0.39	0.24	0.24	0.13	0.39	0.02	0.35	0.34	0.49	0.89	0.49	0.79	0.8	0.73
24	50	0.1	-0.05	0.36	0.27	0.27	0.18	0.31	-0.02	0.24	0.25	0.37	0.69	0.28	0.55	0.58	0.59
24	50	0.3	-0.17	0.32	0.29	0.28	0.18	0.31	-0.08	0.18	0.20	0.32	0.60	0.13	0.35	0.41	0.51
60	20	0.001	0.00	0.44	0.27	0.27	0.04	0.37	-0.07	0.35	0.36	0.57	0.86	0.41	0.8	0.77	0.84
60	20	0.01	-0.01	0.45	0.27	0.27	0.06	0.36	-0.10	0.34	0.34	0.55	0.82	0.36	0.76	0.74	0.81
60	20	0.1	-0.06	0.44	0.27	0.27	0.10	0.32	-0.15	0.28	0.29	0.47	0.67	0.16	0.58	0.58	0.71
60	20	0.3	-0.18	0.41	0.28	0.29	0.11	0.30	-0.20	0.21	0.21	0.40	0.59	0.01	0.42	0.42	0.64
60	50	0.001	0.00	0.45	0.27	0.27	0.01	0.41	-0.04	0.37	0.38	0.62	0.9	0.45	0.81	0.80	0.88
60	50	0.01	0.00	0.45	0.27	0.27	0.02	0.38	-0.06	0.35	0.35	0.61	0.89	0.44	0.81	0.79	0.87
60	50	0.1	-0.04	0.45	0.29	0.29	0.11	0.31	-0.15	0.26	0.26	0.45	0.67	0.18	0.57	0.56	0.70
60	50	0.3	-0.17	0.43	0.30	0.30	0.12	0.30	-0.22	0.19	0.19	0.40	0.58	-0.02	0.38	0.39	0.63

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 16: Estimates of the intracluster correlation coefficient and residual variance based on methods to handle 25% and 50% missing data in categorical cluster level covariate

No. clusters	Cluster size	ICC	MM	MD	SL	FE	AG	σ^2	MM	MD	SL	FE	AG
25% missing													
24	20	0.001	0.00	0.05	0.02	0.03	0.01	0.749	0.74	0.75	0.73	0.73	0.74
24	20	0.01	0.01	0.06	0.04	0.04	0.02	0.74	0.74	0.74	0.73	0.73	0.74
24	20	0.1	0.08	0.18	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
24	20	0.3	0.22	0.40	0.35	0.35	0.34	0.45	0.45	0.45	0.44	0.44	0.45
24	50	0.001	0.00	0.03	0.02	0.02	0.00	0.749	0.75	0.75	0.74	0.74	0.75
24	50	0.01	0.01	0.04	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
24	50	0.1	0.08	0.17	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
24	50	0.3	0.23	0.39	0.35	0.35	0.33	0.45	0.45	0.45	0.44	0.44	0.45
60	20	0.001	0.00	0.04	0.02	0.02	0.00	0.749	0.75	0.75	0.74	0.74	0.75
60	20	0.01	0.01	0.05	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
60	20	0.1	0.08	0.17	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
60	20	0.3	0.22	0.39	0.35	0.35	0.32	0.45	0.45	0.45	0.44	0.44	0.45
60	50	0.001	0.00	0.03	0.01	0.01	0.00	0.749	0.75	0.75	0.74	0.74	0.75
60	50	0.01	0.01	0.05	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
60	50	0.1	0.08	0.17	0.14	0.14	0.10	0.65	0.65	0.65	0.64	0.64	0.65
60	50	0.3	0.22	0.39	0.35	0.35	0.31	0.45	0.45	0.45	0.44	0.44	0.45
50% missing													
24	20	0.001	0.01	0.13	0.08	0.09	0.07	0.749	0.75	0.75	0.71	0.71	0.75
24	20	0.01	0.01	0.14	0.09	0.10	0.08	0.74	0.74	0.74	0.70	0.70	0.74
24	20	0.1	0.07	0.22	0.17	0.19	0.18	0.65	0.65	0.65	0.62	0.62	0.65
24	20	0.3	0.19	0.41	0.35	0.37	0.38	0.45	0.45	0.45	0.43	0.43	0.45
24	50	0.001	0.00	0.13	0.08	0.09	0.06	0.749	0.75	0.75	0.71	0.70	0.75
24	50	0.01	0.01	0.14	0.08	0.10	0.07	0.74	0.74	0.74	0.70	0.70	0.74
24	50	0.1	0.07	0.22	0.17	0.19	0.18	0.65	0.65	0.65	0.62	0.62	0.65
24	50	0.3	0.20	0.41	0.35	0.37	0.38	0.45	0.45	0.45	0.43	0.43	0.45
60	20	0.001	0.00	0.13	0.09	0.09	0.03	0.749	0.75	0.75	0.71	0.72	0.75
60	20	0.01	0.01	0.14	0.10	0.10	0.04	0.74	0.74	0.74	0.71	0.71	0.74
60	20	0.1	0.07	0.23	0.18	0.19	0.14	0.65	0.65	0.65	0.62	0.62	0.65
60	20	0.3	0.20	0.40	0.36	0.37	0.34	0.45	0.45	0.45	0.43	0.43	0.45
60	50	0.001	0.00	0.13	0.08	0.09	0.02	0.749	0.75	0.75	0.71	0.71	0.75
60	50	0.01	0.01	0.14	0.09	0.10	0.03	0.74	0.74	0.74	0.70	0.70	0.74
60	50	0.1	0.07	0.22	0.18	0.18	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	50	0.3	0.20	0.41	0.37	0.37	0.35	0.45	0.45	0.45	0.43	0.43	0.45

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 17: Coverage of true values by the 95% confidence interval of regression coefficients based on methods to handle 25% and 50% missing data in categorical cluster level covariate

No. clusters	Cluster size	ICC	$\beta_0 = 0$ MM	MD	SL	FE	AG	$\beta_1 = 0.5$ MM	MD	SL	FE	AG	$\beta_2 = 0.9$ MM	MD	SL	FE	AG
25% missing																	
24	20	0.001	95	82	96	96	96	95	93	97	97	96	96	93	98	97	96
24	20	0.01	94	82	95	95	96	94	94	97	97	96	95	92	98	97	96
24	20	0.1	92	75	88	89	94	90	88	94	93	94	89	84	90	90	96
24	20	0.3	89	72	74	75	92	91	83	84	85	93	84	77	59	65	95
24	50	0.001	94	85	97	96	95	95	96	98	97	96	95	95	97	97	94
24	50	0.01	92	84	96	95	94	93	94	97	97	95	92	93	96	96	94
24	50	0.1	89	75	83	82	94	92	89	90	87	94	88	85	74	77	95
24	50	0.3	88	69	61	64	91	89	83	67	66	94	85	77	32	38	94
60	20	0.001	96	68	96	96	96	96	81	98	98	96	92	72	97	97	95
60	20	0.01	94	66	94	94	95	94	83	98	98	95	91	73	97	97	94
60	20	0.1	93	56	81	80	95	93	67	94	93	94	79	56	72	73	95
60	20	0.3	92	49	48	50	92	88	57	64	65	92	71	47	23	22	93
60	50	0.001	95	72	97	96	95	95	87	98	98	95	95	81	98	98	95
60	50	0.01	93	68	95	94	94	93	85	97	97	94	93	78	96	96	94
60	50	0.1	93	57	68	68	95	92	70	88	86	93	81	60	47	48	95
60	50	0.3	92	49	31	32	92	88	56	41	40	91	74	47	3	4	93
50% missing																	
24	20	0.001	95	16	34	36	76	86	44	90	82	96	91	81	93	90	93
24	20	0.01	94	16	34	38	74	85	42	89	82	97	84	74	91	85	94
24	20	0.1	88	25	48	49	74	78	47	80	73	95	68	59	72	70	91
24	20	0.3	78	34	56	61	83	82	53	66	68	96	74	58	51	57	92
24	50	0.001	95	12	10	17	74	85	42	84	66	96	96	92	92	83	93
24	50	0.01	92	12	15	20	77	80	43	79	63	97	91	89	89	81	93
24	50	0.1	88	23	33	37	76	77	44	53	53	95	71	62	54	55	91
24	50	0.3	82	34	47	51	81	83	52	49	51	96	73	57	37	41	92
60	20	0.001	95	4	3	2	86	66	4	73	73	74	89	60	91	85	94
60	20	0.01	94	4	3	4	81	62	4	68	67	78	83	49	85	80	91
60	20	0.1	86	8	14	14	79	71	7	58	56	87	66	25	59	53	86
60	20	0.3	69	14	27	27	77	80	12	46	43	87	68	23	37	32	86
60	50	0.001	96	2	0	0	97	58	2	61	59	46	94	66	87	79	98
60	50	0.01	92	2	0	0	94	51	3	49	51	59	91	67	84	77	97
60	50	0.1	90	8	2	3	74	63	8	29	32	86	65	27	37	37	84
60	50	0.3	69	13	14	16	78	81	12	32	27	87	67	20	21	16	85

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

4.5 Discussion

We performed a simulation study to evaluate strategies to handle missing cluster level covariates under the MAR assumption. We varied the total number of clusters, cluster size, ICC, percentage of missingness, and examined the methods when the cluster level covariate was continuous and categorical. We evaluated bias among the fixed and random parameter estimates, as well as coverage of the true values of the fixed parameters.

For both continuous and categorical cluster level covariates, the mixed model produced better estimates of the regression coefficients and ICC when the ICC was low (ICC $\leq$ 0.1) and with 25% missing cluster level data. However, when the ICC was higher (ICC $>$ 0.1), the mixed model generated severely biased estimates of the regression coefficients, which became worse when the percentage of missing data increased to 50%. MI aggregate imputation performed best when the ICC was higher (ICC $>$ 0.1), though it tended to overestimate the ICC similar to the other imputation approaches. When the cluster level covariate was categorical, none of the strategies performed well when 50% of the data were missing. Overall, the worst methods for the missing continuous and categorical cluster level covariate were mean substitution and mode substitution, respectively. Although these single imputation methods are simple, they should not be used, because they produced biased fixed and random effect estimates, especially when the cluster level covariate was a categorical variable.

In general, we found that neither single level MI nor fixed effects MI performed well when imputing cluster level variables, particularly when the percentage of missing data was higher. Both methods produced biased fixed and random parameter estimates, and were undercoveraged for the fixed parameters. Van Buuren studied missing data strategies to handle missing outcomes and covariates at the individual level under MAR. He compared the mixed model, single level MI ignoring clustering, fixed effects MI, and multilevel MI, which incorporates clustering into the imputation process via the Gibbs sampler. He found fixed effects MI to perform reasonably well, and found single level MI ignoring clustering to be less successful in generating appropriate fixed and random parameter estimates [31]. Single level MI is still generally used in practice [39], even though it has been shown to perform well only under the restrictive assumption that the continuous individual level outcome is MCAR and the ICC is very low (ICC $<$ 0.005) [30]. Single level MI should not be used to impute cluster level covariates because it has been shown to perform poorly, regardless of whether the data are MCAR [70, 71] or MAR.

A strength of our study is that we assessed missing data strategies under the MAR assumption, which may be a more likely scenario in practice. Along with the continuous cluster level covariate, we also investigated the behavior of methods when the cluster level covariate was categorical, which may be more widely found among cluster level data. We evaluated scenarios with small and large total number of clusters, cluster sizes, and ICC.

We did not examine the scenario in which missing data occur in multiple levels within the hierarchical structure. For example, along with missing cluster level covariates, missing data can occur among individual level outcomes and covariates simultaneously. This scenario is perhaps most commonly seen in practice. Van Buuren studied the scenario in which missing data occur among individual level outcomes and covariates simultaneously, and found multilevel MI to perform best, though still not ideal [31]. Another, more complex setting to examine is a three-level model, such as a longitudinal cluster randomized trial, which includes clusters, individuals per cluster, and measurements per individual. Further investigation of appropriate approaches when missing data occur in different places among clustered data is needed. Although it is possible to test between MCAR and MAR, it is not possible to test between MAR and MNAR since the data are missing. For this reason, it would be beneficial to examine the performance of approaches under departures from the MAR assumption in future.

Based on our simulations, we recommend using the mixed model for missing cluster level covariates when the ICC is small ($\rho \leq 0.1$) and the percentage of missing data is low ($\leq 25\%$), as long as there are a large number of clusters. Otherwise, MI aggregate imputation should be used to impute missing cluster level covariates, though caution should be taken if the percentage of missing cluster level covariates is high. Mean and mode substitution are not recommended as effective strategies to imputed missing cluster level covariates.

5 CONCLUSIONS AND FUTURE WORK

This dissertation explored statistical approaches related to handling missing data in CRTs. I found that the majority of recently published CRTs reported some missing outcome data in their primary analysis. These findings show that missing data is a common problem in CRTs, and should be accommodated with appropriate statistical methods. I also found that a few CRTs are reporting a sensitivity analysis for missing data, with only a fraction of them weakening the missingness assumption from the primary analysis.

In order to promote carrying out a sensitivity analysis for missing data in CRTs, I proposed an approach using the PMM framework to deal with missing data under the MNAR assumption. I implemented multilevel MI and transformed imputed MAR values by multiplying them by a sensitivity parameter, k, to create imputed MNAR values. By doing this, researchers can evaluate whether results due to differing missing data assumptions are reasonable for their CRT.

Lastly, I investigated strategies for handling missing cluster level covariates under MAR, including the linear mixed effects model, single imputation, single level MI ignoring clustering, MI incorporating clusters as fixed effects, and MI at the cluster level using aggregated data. I found that the linear mixed effects model performed well when the ICC was low (ICC ≤ 0.1) and the percentage of missing data was low ($\leq 25\%$). When the ICC was higher, (ICC > 0.1) MI at the cluster level using aggregated data performed best with 25% missing data. None of the missing data strategies performed particularly well with 50% missing cluster level covariates.

In the future, I plan to continue my work in missing data and CRTs. The PMM I proposed to handle MNAR data in CRTs dealt with the simplest case of two time points, baseline and follow-up. Assuming that all individuals were measured at baseline, there were only two missing data patterns: those who were also measured at follow-up (non-responders) and those who dropped out after baseline (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. Therefore, I plan to extend the PMM approach for CRTs to handle multiple time points with missing data. I plan to approach this scenario by first performing multilevel MI to impute missing outcome observations under the MAR assumption. Imputations for missing observations will be conditional on previous and subsequent measurements in order to utilize all informational available. MAR imputed values will be multiplied by a specified k in order to transform them into MNAR imputed values.

Another approach for the PMM is to use missing value restrictions, which imputes missing data based on other missing data patterns. This technique has been extensively discussed for longitudinal data, but has not yet been implemented when there is an added level of

clusters. Little (1993) proposed three strategies to identify unknown parameters: complete case missing value (CCMV), neighboring case missing value (NCMV), and available case missing value (ACMV) [41]. CCMV uses data from those who completed the trial to impute missing observations in other patterns. NCMV imputes missing observations by using data from the next identified dropout pattern up. ACMV imputes missing observations by using available data from individuals in higher identified patterns, which is equivalent to assuming MAR. To extend these strategies in CRTs, multilevel MI can be used to impute missing observations conditional on other dropout patterns. In the case of three measurements $y = (y_1, y_2, y_3)^T$, we define three dropout patterns: pattern 3 contains individuals who completed the trial, pattern 2 contains the individuals with missing y_3, and pattern 1 contains the individuals with missing y_2 and y_3. Under CCMV all missing observations are imputed conditional on the observed data in pattern 3. Under NCMV, missing y_3 for individuals in pattern 2 are imputed conditional on data from pattern 3, the next pattern up. For individuals in pattern 1, missing y_2 and y_3 are imputed conditional on pattern 2.

In my simulation study, my PMM models were specified in concordance with the generated data. In the future, I plan to assess consequences of misspecifying the PMM model in CRTs since it is not known whether the model is correctly specified in most circumstances. I plan to do this by generating data from a true model that is different from the model used for analysis, and evaluating how this affects parameters of interest and their corresponding standard errors.

In my third paper, I examined the scenario of a single missing cluster level covariate in multilevel data. In CRTs, missing data can occur in multiple roles simultaneously. For example, along with missing cluster level covariates, missing data can also occur among individual level outcomes and covariates. Since this scenario is perhaps more commonly encountered among CRTs, I plan to investigate the performance of strategies when missing data are present in multiple levels within the hierarchical structure. This will be done by first generating multilevel data with both individual level and cluster level variables. Missing data will be created in the cluster level covariate, individual level covariate, and outcome under MAR. I will then compare the performance of missing data strategies such as the mixed model, single level MI, fixed effects MI, and multilevel MI with respect to bias, standard errors, and coverage of parameters of interest.

PRINCIPAL ABBREVIATIONS

CRT	Cluster Randomized Trial
CV	Coefficient of Variation
GEE	Generalized Estimating Equation
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
ICC	Intracluster Correlation Coefficient
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
MI	Multiple Imputation
MNAR	Missing Not At Random
PMM	Pattern Mixture Model

APPENDIX A: MANUSCRIPT 1

Open Access Protocol

BMJ Open

Statistical analysis and handling of missing data in cluster randomised trials: protocol for a systematic review

Mallorie Fiero, Shuang Huang, Melanie L Bell

To cite: Fiero M, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomised trials: protocol for a systematic review. *BMJ Open* 2015;**5**: e007378. doi:10.1136/bmjopen-2014-007378

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2014-007378).

Received 4 December 2014
Revised 31 March 2015
Accepted 9 April 2015

Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, USA

Correspondence to
Mallorie Fiero;
mfiero@email.arizona.edu

ABSTRACT

Introduction: Cluster randomised trials (CRTs) randomise participants in groups, rather than as individuals, and are key tools used to assess interventions in health research where treatment contamination is likely or if individual randomisation is not feasible. Missing outcome data can reduce power in trials, including in CRTs, and is a potential source of bias. The current review focuses on evaluating methods used in statistical analysis and handling of missing data with respect to the primary outcome in CRTs.

Methods and analysis: We will search for CRTs published between August 2013 and July 2014 using PubMed, Web of Science and PsycINFO. We will identify relevant studies by screening titles and abstracts, and examining full-text articles based on our predefined study inclusion criteria. 86 studies will be randomly chosen to be included in our review. Two independent reviewers will collect data from each study using a standardised, prepiloted data extraction template. Our findings will be summarised and presented using descriptive statistics.

Ethics and dissemination: This methodological systematic review does not need ethical approval because there are no data used in our study that are linked to individual patient data. After completion of this systematic review, data will be immediately analysed, and findings will be disseminated through a peer-reviewed publication and conference presentation.

Strengths and limitations of this study

- To our knowledge, this is the first systematic review to evaluate statistical analysis and handling of missing outcome data in cluster randomised trials (CRTs).
- The study uses prespecified search strategy, study selection criteria and data extraction strategy, which minimises the potential for bias during the review process.
- Study selection criteria encompass a wide range of CRTs including stepped wedge designs and feasibility studies.
- Pilot testing will be performed on several trials by three independent reviewers. Data collection will be carried out by two independent reviewers to ensure accuracy.
- The study is subject to potential selection bias. Researchers who include terms such as 'cluster randomised' in the title or abstract may be more likely to follow the CONSORT statement compared with trials that do not include these terms. Researchers who do not realise their trials are CRTs are likely to use less robust methods.

INTRODUCTION

Cluster randomised trials (CRTs) randomise groups of participants to intervention arms, as opposed to individual participants. CRTs are frequently used in health research to minimise intervention arm contamination, or to assess interventions that can only be carried out at a cluster (eg, physician, centre) level.[1 2]

Cluster-level allocation generates several issues for statistical analysis. Participants cannot be assumed to be independent because of the similarity among participants within the same cluster. The intracluster correlation coefficient (ICC) is the statistical measure of this within-cluster dependence. Suppose some variable y was measured on n individuals divided into k clusters. The ICC, ρ, is the proportion of variance due to clustering, given by:

$$\rho = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_e^2}$$

where σ_k^2 and σ_e^2 denote the between-cluster and within-cluster variances, respectively. Ignoring clusters in the analysis can lead to falsely low p values, overly narrow CIs, and increased type I error rates.[3 4]

Missing data lead to a reduction of power, compromise the benefits of randomisation and are a potential source of bias. In practice, there will almost always be some missing data.[5 6] Recent reviews in individual randomised trials have found that the majority have missing outcome data.[7–10] Missing data mechanisms have been broadly categorised

into the following three classes. Data are said to be missing completely at random (MCAR) if the reason for a missing observation is unrelated to observed values of the outcome and covariates. MCAR is a strong assumption and unlikely in most trials. A more reasonable assumption is missing at random (MAR), where missingness does not depend on the unobserved data, conditional on the observed data. Lastly, data are considered missing not at random if missingness depends on the unseen value of that observation even after conditioning on fully observed data.[6 11]

Several reviews have been published regarding CRTs.[12–22] Most have reported inadequate accounting for clustering in sample size and analysis. One review of CRTs published in 2011 focused on imputation techniques with respect to handling missing data and did not discern between missing covariates or outcomes.[23] Modelling approaches can differ based on whether outcomes or covariates are missing: if covariates are missing, multiple imputation (MI) or an unadjusted model can be used. If outcomes are missing, maximum likelihood estimation using mixed models, for example, can provide unbiased estimation in certain cases (see below). Additionally, there was no distinction of whether trials used a complete case analysis, generalised estimating equations (GEE) or mixed models with respect to handling missing data in the primary analysis. Distinguishing between these methods is important, as they may provide valid estimates under certain missing data assumptions. Our objective is to provide a comprehensive review of analytical approaches for handling missing outcome data in CRTs. The primary aims of our review are to evaluate approaches used to analyse primary outcome data in CRTs and investigate methods used to handle missing outcome data in primary and sensitivity analysis. As a secondary aim, we will evaluate methods for achieving balance in CRTs by examining the proportions of CRTs that use stratification, matching or minimisation.

METHODS

Our systematic review will investigate statistical analyses and missing data strategies used in CRTs. This section contains an introduction of commonly used statistical approaches and missing data methods used for analysing clustered data, followed by a detailed description of our methodological strategy based on guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement.[24]

Statistical approaches for analysing CRTs

Two standard approaches to analyse CRTs include analysis at the cluster level and analysis at the individual level. Cluster-level analysis involves reducing all observations within a cluster to a single summary measure, such as a cluster mean or proportion. Standard statistical tests (eg, t tests, linear regression models) can then be performed since each data point can now be considered independent.[4 25] Even though cluster-level analysis solves the problem of dependent data, reducing observations to single summary statistics leads to a reduction in sample size and as a result, statistical power. Modelling techniques incorporating individual-level covariates in cluster-level analysis, such as generalised linear mixed models (GLMM) and GEE, have also been developed.[26 27] GEE and GLMM explicitly involve intracluster correlation in the modelling process, which enables a more realistic model of the clustered data. An advantage of these types of models is the ability to control for confounding at the individual level and reduce bias. However, drawbacks of this approach are that they are more computationally intensive and require a higher sample size of relatively large clusters.[25 28]

Missing data methods in CRTs

Common approaches for handling missing outcome data include complete case analysis, single imputation, MI and model-based analysis. Complete case analysis excludes participants with missing data and is valid (produces unbiased estimates) if missingness is independent of the outcome, given covariates.[29] Single imputation strategies fill-in missing data with a single value, thereby underestimating uncertainty. Under the MAR assumption, MI takes into account uncertainty by replacing each missing value with a set of possible values to create multiple imputed data sets. However, most implementations are single level, ignoring the hierarchical data structure of CRTs. Multilevel MI reflects the lack of independence found within clusters due to the multilevel structure of CRTs.[30 31] Model-based methods include linear mixed models, valid for MAR data, if the model is specified correctly, and GEE, which is valid under the stronger MCAR assumption as long as there are a large number of clusters.[28 32] Inverse probability weighting (IPW) is used to make a valid complete case analysis under MAR by weighting complete cases with the inverse of their probability of having data observed.[33] The IPW approach is relatively simple to carry out when missing values have a monotone pattern and can be applied to GEE. However, there is possible instability when weights are extremely large, which can lead to biased estimates and high variance in small samples.[6]

Search strategy

CRTs published in English between August 2013 and July 2014 will be sought. Two authors (MF and SH) will systematically search for CRTs indexed in the following electronic bibliographic databases: PubMed, Web of Science (all databases) and PsycINFO. The search strategy will include the terms "cluster randomized [randomised]", "cluster and trial", "community trial", "community randomized [randomised]" or "group randomized [randomised]" found in titles and abstracts. An example of our search strategy including search terms is found in online supplementary file 1.

Inclusion and exclusion criteria

We will include all CRT designs, including stepped wedge trials.[34] We will exclude protocols of trials, observational studies, secondary reports of trials, studies in which no data were collected at the individual level and quasi-experimental cluster designs. Trials with survival outcomes will also be excluded, as missing time-to-event data are handled quite differently to other types of outcome data.

Study selection

Two independent reviewers (MF and SH) will identify eligible studies using the search strategy. All studies will be imported using EndNote (EndNote X6, Thomson Reuters, New York, USA). The reviewers will remove duplicates and go through titles and abstracts to identify eligible studies. Full-text articles will be retrieved if the reviewer identified the article to answer 'yes' or 'unclear' to all selection criteria. The reviewers will collect and evaluate the full text article, and identify relevant studies based on study inclusion criteria. Reviewers will keep track of the number of studies excluded from each screening step.

Sample size

We hypothesise 90% of trials having some missing outcome data. We estimate that a sample size of 86 papers will result in a margin of error of 6 percentage points (95% CI 84 to 96).

Data extraction strategy

Pilot testing of coding will be carried out with both reviewers (MF and SH) and the senior author (MLB). All piloted papers will be included in the review. Two independent reviewers (MF and SH) will collect data from each study using a standardised, prepiloted data extraction template. Disagreements over the eligibility or data extraction of particular studies will be handled by consensus or a third reviewer where consensus was not achieved.

Extracted information will include: general information (journal, author, date of publication, pilot/feasibility study or stepped wedge); characteristics of the primary outcome (type of outcome, how often outcome was collected, how outcome was treated in the primary analysis); characteristics of study participants (unit or randomisation, stratification/matching/minimisation used, number of clusters randomised, total number of participants randomised, response rate at time period of primary analysis, if survey data); details of sample size calculation (accounted for clustering in calculation, reported ICC or coefficient of variation (CV), accounted for missing outcome data in calculation, reported attrition rate in sample size calculation); primary analysis (statistical method used in primary analysis, adjustment (unadjusted, adjusted for design variables such as stratification, adjusted beyond stratification variables), clustering accounted for in analysis, observed ICC or CV, GEE correction type); information on missing data (number (and proportion) of clusters with missing outcome, number (and proportion) of participants with missing outcome, reasons for missing data, method to handle missing data in primary analysis and sensitivity analysis). If any of the items were unclear, including the amount of missing data and method used to handle missing data, we specified it as 'unclear'. Specific details on data items, including relevant coding used during the data extraction process and definitions, are given in online supplementary file 2.

Method of analysis

Our analysis strategy follows closely after reviews by Wood *et al*[7] and Bell *et al*,[10] which both assessed missing outcomes in individually randomised trials. We will present a synthesis of the findings by first describing characteristics of the primary outcome and study participants of the included studies. We will then calculate the proportion of trials reporting some missing data at the individual and cluster level. This will be determined from flow diagrams or text with respect to follow-up of clusters and individuals. Of those who reported some missing data, we will calculate the proportion of trials that carried out complete case analysis, single imputation, MI, GEE or a mixed model to handle missing data in the primary analysis. Similar computations for trials that report sensitivity analysis for missing data will also be performed. We will quantify the number of trials that weakened the missingness assumption of their primary analysis to perform their sensitivity analysis as suggested by the Panel on Handling Missing Data in Clinical Trials, recently commissioned by the National Research Council.[6]

To evaluate prevention and planning, we will record whether sample size calculations were reported and if trials accounted for clustering and missing data. We will describe the details of analysis of primary outcomes and compare observed versus expected attrition rates and ICCs (or CVs). Quality of trials will not be assessed.

DISCUSSION

To our knowledge, this is the first systematic review to evaluate statistical analysis and handling of missing outcome data in CRTs. We have a prespecified search strategy, study selection criteria and data extraction strategy. Systematic reviews are complicated and require judgements that should not rely on conclusions of the studies included in the review.[35] By predefining our methodology, we are minimising the potential for bias during the review process. Additionally, our study selection criteria encompass a wide range of CRTs including stepped wedge designs and feasibility studies. Pilot testing will be performed on several trials by three independent reviewers. Data collection will be carried out by two independent reviewers to ensure accuracy.

A limitation of this systematic review is the difficulty in identifying CRTs since many do not use the term 'cluster' in the title or abstract. In an effort to alleviate this issue, we will use other commonly used terms for cluster randomisation including 'community randomised' or 'group

randomised'. This allows us to reach a wider range of trials that may have been missed otherwise.

Furthermore, our systematic review is subject to potential selection bias. Researchers who include terms such as 'cluster randomised' in the title or abstract may be more likely to follow the CONSORT statement compared with trials that do not include these terms.[36] Researchers who do not realise their trials are CRTs are likely to use less robust methods.

Language bias may be introduced since we have limited our search to CRTs published in the English language.

Including studies with survival outcomes may influence missing data rates since participants are censored at dropout. We did not consider CRTs of which the primary outcome was survival because different statistical issues arise in comparison to trials with non-survival outcomes.

This review will allow us to examine current statistical methods used in practice with respect to missing outcomes in CRTs. Based on our results, we will be able to make recommendations for areas where reporting and conduct may need improvement.

Contributors MF and MLB conceptualised the study. MF drafted the manuscript and incorporated comments from authors for successive drafts. SH and MLB contributed to design and content. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement After completion of this systematic review, data will be immediately analysed and findings will be disseminated through a peer-reviewed publication and conference presentations.

REFERENCES

1. Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. London Arnold Publishers, 2000.
2. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ* 1998;317:1171–2.
3. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;108:100–2.
4. Campbell MK, Mollison J, Steen N, *et al*. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract* 2000;17:192–6.
5. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Stat Methods Med Res* 2014;23:440–59.
6. National Research Council. The prevention and treatment of missing data in clinical trials. In: *Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. Washington DC: National Academies Press, 2010.
7. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004;1:368–76.
8. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? *Clin Trials* 2007;4:350–6.
9. Fielding S, Maclennan G, Cook JA, *et al*. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* 2008;9:51.
10. Bell ML, Fiero M, Horton NJ, *et al*. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol* 2014;14:118.
11. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
12. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol* 1990;19:795–800.
13. Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;85:1378–83.
14. Smith PJ, Moffatt ME, Gelskey SC, *et al*. Are community health interventions evaluated appropriately? A review of six journals. *J Clin Epidemiol* 1997;50:137–46.
15. Chuang JH, Hripcsak G, Jenders RA. Considering clustering: a methodological review of clinical decision support system studies. *Proc AMIA Symp* 2000:146–50.
16. Hayes RJ, Alexander ND, Bennett S, *et al*. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Methods Med Res* 2000;9:95–116.
17. Isaakidis P, Ioannidis JP. Evaluation of cluster randomized controlled trials in sub-Saharan Africa. *Am J Epidemiol* 2003;158:921–6.
18. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ* 2003;327:785–9.
19. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4:21.
20. Eldridge S, Ashby D, Bennett C, *et al*. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ* 2008;336:876–80.
21. Eldridge SM, Ashby D, Feder GS, *et al*. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004;1:80–90.
22. Varnell SP, Murray DM, Janega JB, *et al*. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health* 2004;94:393–9.
23. Diaz-Ordaz K, Kenward MG, Cohen A, *et al*. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials* 2014;11:590–600.
24. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
25. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med* 2002;9:330–41.
26. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;42:121–30.
27. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. John Wiley & Sons, 2012.
28. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med* 2007;26:2–19.
29. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920–31.
30. Van Buuren S. Multiple imputation of multilevel data. In: Hox JJ, Roberts JK, eds. *Handbook of advanced multilevel analysis*. Psychology Press, 2011:173–96.
31. Caille A, Leyrat C, Giraudeau B. A comparison of imputation strategies in cluster randomized trials with missing binary outcomes. *Stat Methods Med Res* 2014 Published Online First 7 Apr 2014. doi:10.1177/0962280214530030
32. Robins J, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995;90:106–21.
33. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994;89:846–66.
34. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28:182–91.
35. Higgins JP, Green S. *Cochrane handbook for systematic reviews of interventions*. Wiley Online Library, 2008.
36. Campbell MK, Elbourne DR, Altman DG, *et al*. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328:702–8.

APPENDIX B: MANUSCRIPT 2

Fiero et al. Trials (2016) 17:72
DOI 10.1186/s13063-016-1201-z

Trials
65

RESEARCH Open Access

Statistical analysis and handling of missing data in cluster randomized trials: a systematic review

Mallorie H. Fiero*, Shuang Huang, Eyal Oren and Melanie L. Bell

Abstract

Background: Cluster randomized trials (CRTs) randomize participants in groups, rather than as individuals and are key tools used to assess interventions in health research where treatment contamination is likely or if individual randomization is not feasible. Two potential major pitfalls exist regarding CRTs, namely handling missing data and not accounting for clustering in the primary analysis. The aim of this review was to evaluate approaches for handling missing data and statistical analysis with respect to the primary outcome in CRTs.

Methods: We systematically searched for CRTs published between August 2013 and July 2014 using PubMed, Web of Science, and PsycINFO. For each trial, two independent reviewers assessed the extent of the missing data and method(s) used for handling missing data in the primary and sensitivity analyses. We evaluated the primary analysis and determined whether it was at the cluster or individual level.

Results: Of the 86 included CRTs, 80 (93 %) trials reported some missing outcome data. Of those reporting missing data, the median percent of individuals with a missing outcome was 19 % (range 0.5 to 90 %). The most common way to handle missing data in the primary analysis was complete case analysis (44, 55 %), whereas 18 (22 %) used mixed models, six (8 %) used single imputation, four (5 %) used unweighted generalized estimating equations, and two (2 %) used multiple imputation. Fourteen (16 %) trials reported a sensitivity analysis for missing data, but most assumed the same missing data mechanism as in the primary analysis. Overall, 67 (78 %) trials accounted for clustering in the primary analysis.

Conclusions: High rates of missing outcome data are present in the majority of CRTs, yet handling missing data in practice remains suboptimal. Researchers and applied statisticians should carry out appropriate missing data methods, which are valid under plausible assumptions in order to increase statistical power in trials and reduce the possibility of bias. Sensitivity analysis should be performed, with weakened assumptions regarding the missing data mechanism to explore the robustness of results reported in the primary analysis.

Keywords: Cluster randomized trials, Missing data, Dropout, Sensitivity analysis

Background

In cluster randomized trials (CRTs), groups of participants, rather than individuals, are randomized to intervention arms. CRTs are often adopted to reduce treatment contamination or if individual randomization is unsuitable and are an increasingly popular approach in comparative effectiveness research [1–4]. In cluster-level allocation, participants cannot be assumed as independent because of the similarity among participants within the same cluster or cluster characteristics, leading to intracluster correlation, or equivalently, between-cluster variation [3]. Two potential pitfalls with respect to CRTs are handling missing data and not accounting for clustering in the primary analysis.

Missing data decreases power and precision and can lead to bias by compromising randomization. For example, treatment arm imbalance with respect to missing data is likely to introduce bias when the outcome is related to the reason for patient withdrawal. Even if missing

* Correspondence: mfiero@email.arizona.edu
Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, 1295 N. Martin Ave., Drachman Hall, P.O. Box 245163, Tucson, Arizona 85724, USA

outcome data are balanced across treatment arms, differing reasons for the missing outcome can cause bias [5]. Reviews of individually randomized controlled trials have discovered that most trials have some missing outcome data [6, 7]. Few reports have discussed missing data in CRTs, despite its high likelihood and the recognition that it poses a serious threat to research validity, as discussed by the National Research Council and the Patient Centered Outcomes Research Institute [5, 8].

Missing data mechanisms are commonly classified into the following three categories. Data are considered to be missing completely at random (MCAR) if missingness is independent of the observed outcomes and covariates. MCAR is a strong assumption and is not likely in most clinical trials. A more sensible assumption is missing at random (MAR), where missingness does not depend on unobserved data after conditioning on the observed data. Data are termed missing not at random (MNAR) if missingness is dependent on unobserved data values even after conditioning on fully observed data [9, 10].

The most common approach for handling missing outcome data is a complete case analysis, which excludes individuals with missing data. This approach yields unbiased estimation if missingness is independent of the outcome, given the covariates [11]. Additional approaches include imputation (single and multiple) and model-based methods. Single imputation strategies, such as the popular last observation carried forward (LOCF) used in longitudinal studies, or mean substitution, replaces missing data with a single number, which underestimates uncertainty [12, 13]. LOCF also makes unlikely assumptions about an individual's trajectory and can lead to either under- or overestimation of treatment effects [14].

Under the MAR assumption, multiple imputation (MI) considers uncertainty by filling in missing data from a distribution of likely values. Analysis is performed on each dataset and the results combined using specified algorithms. Most implementations of MI are single level, ignoring the multilevel structure of CRTs. Multilevel MI incorporates the lack of independence found within clusters due to the hierarchical data structure found in CRTs [15].

Likelihood based mixed models are valid for MAR data if the model is specified correctly, while unweighted GEE are valid under MCAR if there are a large number of clusters [16, 17]. In order to make a valid complete case analysis under the MAR assumption, inverse probability weighting (IPW) weights complete cases with the inverse of their probability of being observed [18]. Although IPW is relatively simple to perform with monotone missing data, it is prone to large weights, which cause unstable estimates and high variance [10].

The second difficulty regarding CRTs is accounting for clustering in the primary analysis. Ignoring clustering can lead to confidence intervals that are too narrow and increased type I error rates [19, 20]. In order to account for clustering, analysis can be performed at the cluster level or at the individual level. Cluster-level analysis reduces observations within a cluster to an aggregate value and then analyzes each independent data point [20, 21]. Although cluster level analysis alleviates the issue of dependent data, reducing all observations within a cluster to a single summary measure decreases the sample size and power. Analyses at the individual level using general linear models (GLMs) account for non-independent observations within clusters through robust standard errors or adjust using the design effect, an inflation factor used to achieve the same power of an individually randomized trial [22]. Modeling techniques such as generalized estimating equations (GEE) [23] and mixed models [24] explicitly involve intracluster correlation in the modeling process, which enables a more realistic model of the clustered data [24, 25]. Although these models can reduce bias by controlling for confounding at the individual level, they require a higher sample size of a large number of clusters [1, 17, 21].

There have been several reviews on methodological aspects of CRTs (see for example, Simpson et al. [26] and Campbell et al. [27], and the references therein). Diaz-Ordaz et al. [28] reviewed the imputation methods used to handle missing data in CRTs but did not distinguish whether a complete case analysis, GEE, or mixed model was used to handle missing data in the primary analysis, as these approaches provide valid estimates under differing missing data assumptions. Thus, our objective was to provide a comprehensive review of how missing data are being dealt with in CRTs. The primary aims of our review were to accomplish the following:

1. Identify the proportion of CRTs with missing data at the cluster and individual level.
2. Examine the analytical approaches for the primary analysis to find out whether
   a. whether missing data had been accommodated and
   b. whether clustering had been accounted for.
3. Identify the proportion of CRTs reporting a sensitivity analysis for missing data.

Secondary aims included assessing the techniques for achieving balance in CRTs (stratification, matching, or minimization), the differences between observed and expected attrition rates, and the intracluster correlation.

Methods

This study was a systematic review of a sample of CRTs published between August 2013 and July 2014. Our methodological strategy was based on guidelines from the Preferred Reporting Items for Systematic Reviews and

Meta-Analysis (PRISMA) statement (See Additional file 1 for compliance details) [29]. We have reported a detailed protocol for this study elsewhere [30].

Eligibility criteria

Eligible studies were restricted to CRTs published in English between August 2013 and July 2014. We included all types of CRTs with human participants, including stepped wedge trials that were reported in the databases listed below [31, 32]. We excluded trial protocols, non- or quasi-experimental designs, secondary trial reports, cost-effectiveness reports, and studies where no individual-level data were collected. We also excluded trials where the primary outcome was survival, as time-to-event analyses handle censored data differently than other types of data.

Literature search and study selection

Two authors (MF and SH) electronically searched for studies found in PubMed, Web of Science (all databases), and PsycINFO. Titles and abstracts were searched containing the terms "cluster randomized [randomised]," cluster and trial, "community trial," "community randomized [randomised]," or "group randomized [randomised]." Two independent reviewers (MF and SH) screened titles and abstracts, removed duplicates, and screened full texts.

Both reviewers (MF and SH) and the senior author (MB) performed pilot testing of the data extraction form. All papers used for piloting were included in the systematic review. The reviewers extracted data from each trial using a standardized, pilot-tested form. Disagreements over study eligibility or data extraction were resolved by discussion or with the assistance of a third reviewer (MB) when needed.

Sample Size

Based on previous literature, it was estimated that about 90 % of trials would report some missing outcome data [6, 7]. Using the formula for a 95 % confidence interval (CI) for a proportion, we estimated that a sample size of 86 papers would result in an acceptable 95 % CI for the hypothesized 90 % of studies having some missing outcome data (95 % CI of 84 to 96).

Analysis

We defined the number of clusters (and participants) in each trial as the number of clusters (and participants) at randomization. We computed the average number of participants per cluster by dividing the number of participants by the number of clusters.

Description and handling of missing data

We evaluated the degree of missing data and the method(s) for handling missing data in the primary analysis for each trial. The primary analysis was defined as the main analysis of the primary outcome. When multiple primary outcomes were reported, we used the first outcome listed in the methods section. For primary outcomes measured repeatedly, we used the final follow-up time point to calculate the missing proportion, unless a different time point was specified for the primary analysis.

The proportion of clusters with a missing outcome was calculated as the number of entire clusters with a missing outcome (generally due to the entire cluster dropping out) divided by the number of clusters randomized. Clusters that were randomized but failed to recruit were considered missing. A similar calculation was carried out for the proportion of participants with a missing outcome. In cases where an entire cluster dropped out, the missing data rate was included in our calculation of missing participants. If the trial had longitudinal data, we calculated the missing rate at the last time point or time point of the primary analysis if specified. Of those who reported some missing data, we identified the statistical methods used to handle missing data, classified into the following categories: complete case, single imputation (such as worst case or LOCF), MI (single level or multilevel), GEE, mixed model or IPW. Technically, mixed models and GEE are considered complete case analyses. However, we make the distinction because these are model-based methods. Mixed models are valid under MAR, and GEE can be modified to be valid under MAR. We also reported methods for missing data for trials indicating greater than or less than 10 % missing data at the individual level. We indicated that a trial presented a sample size calculation if there was enough detail for replication. We recorded whether sample size calculations accounted for missing data, and compared observed and expected attrition rates with the mean absolute difference. If a range was reported for attrition rates, we used the upper bound.

Sensitivity analysis for missing data

We computed the number of trials that reported performing a sensitivity analysis and determined the method(s) used to deal with missing data in any sensitivity analysis. Sensitivity analysis was defined as any analysis performed to assess the robustness of the primary results due to changes in assumptions regarding missing outcome data. We also reported methods for sensitivity analysis for trials indicating greater than or less than 10 % missing data at the individual level. We quantified the number of trials that weakened the missingness assumption of their primary analysis (MCAR → MAR → MNAR) to perform their sensitivity analysis as suggested by the Panel on Handling Missing Data in Clinical Trials [10].

Accounting for clustering in the primary analysis

For each trial, we calculated the proportion of CRTs performing an individual-level or cluster-level analysis and

whether the analysis accounted for clustering. Individual level analyses were categorized into the following groups: basic inferential test (such as *t*-test or chi-square)/GLM (such as linear or logistic regression), GEE, or mixed model. The analysis accounted for clustering if the basic inferential test or GLM obtained robust standard errors or was adjusted using the design effect, if GEE introduced an exchangeable correlation structure for clusters, or if the mixed model used clusters as a random effect. Basic inferential tests/GLMs could also be carried out as a cluster-level analysis. We examined whether the primary analysis was unadjusted, adjusted for baseline variables, adjusted for balance variables such as stratification, or adjusted for additional covariates.

The intracluster correlation coefficient (ICC) measures the degree of similarity among responses within a cluster and is defined as the proportion of total variance due to between-cluster variation. The coefficient of variation (CV) is an alternate measure of between-cluster variability and is defined by the ratio of the standard deviation of cluster sizes to the mean cluster size [3]. We recorded whether trials accounted for clustering in sample size calculations and compared the observed and expected ICCs (or CVs) with the mean absolute difference. If a range was reported for the ICC (or CV), we used the upper bound.

Results

We identified 3,674 records through our electronic database search after removing 2,164 duplicates. We screened 1,510 of the remaining records, of which, 1,049 were excluded, based on titles or abstracts, as not meeting our eligibility criteria. We examined the full texts of the remaining 461 trials and excluded a further 59 trials, as they did not meet eligibility criteria. Of the 402 eligible reports, we used six for piloting and randomly selected 80 others, thereby including 86 trials in the analyses (Fig. 1). The full list of the included studies is given in Additional file 2.

Table 1 presents the general characteristics of the included trials. In total, the median number of clusters randomized was 24, with a range of 2 to 1,552. Three trials were unclear in the number of clusters randomized. The median number of individuals included was 688, with a range of 49 to 117,100. The average number of individuals per cluster ranged from 1 to 1,105. Of the 65 trials that collected the outcome repeatedly, 36 (55 %) used all of the information in the primary analysis by treating the outcome as a repeated measurement, while 29 (45 %) were analyzed at a single time point. Forty-four (51 %) trials used balance techniques to ensure balance after randomization. Stratification was the most common method (27, 61 %), a subset of which also used matching (1) and minimization (1). Fourteen (32 %) of the trials carrying out balance methods used matching, and three (7 %) used minimization.

Description and handling of missing data

Twenty-seven (31 %) trials reported having whole clusters missing in the primary analysis (Table 2). Of these, the median amount of clusters missing was 7 %, with a range of 0.8 to 51 %. Three trials had an unclear number of clusters missing. Reasons for whole clusters missing included closures, natural disasters, a lack of eligible participants, and an inability to retrieve data. Figure 2 displays the proportions of included individuals with missing outcomes. Eighty (93 %) trials reported having some missing data at the individual level. Of these trials, the median amount of missing individual level data was 19 %, with a range of 0.5 to 90 %. Eight trials were unclear in the amount of individual-level missing data. Of the trials reporting some missing data, 61 (76 %) reported reasons for individuals missing, two (2 %) reported missing data due to missing covariates in the adjusted analyses, and 17 (22 %) were unclear or did not report reasons for individuals missing.

The most common approach for handling missing data in the primary analysis was a complete case analysis (44, 55 %) (Table 3). Eighteen (22 %) trials used mixed models. Six (8 %) carried out single imputation methods: three used worst-case imputation, two used LOCF, and one used baseline observation carried forward. Four (5 %) trials used unweighted GEE. Two (2 %) trials performed MI, although neither used multilevel methods. A MAR assumption for the primary analysis was made in 20 (25 %) of the trials with missing data.

Of the 58 trials reporting more than 10 % missing data at the individual level, 31 (53 %) used complete case analysis, 17 (29 %) used mixed models, five (9 %) used single imputation, two (3 %) used MI, and three (3 %) used methods that were unclear. Of the 14 trials reporting less than 10 % missing data at the individual level, 10 (71 %) used complete case, three (21 %) used unweighted GEE, and one (7 %) used single imputation.

Sixty (70 %) trials presented a sample size calculation, of which 28 (47 %) accounted for missing data via sample size inflation. Twenty-six of these trials accounted for missing data at the individual level, either by dividing by (1 - the estimated dropout rate) or multiplying by (1 + the estimated dropout rate). Two trials also accounted for missing data at the cluster level by including extra clusters in each trial arm. Two trials mentioned sample size inflation but were unclear if they accounted for missing data at the cluster or individual level. Of the 21 trials that reported an expected and observed attrition rate, one trial estimated a higher attrition rate than observed, whereas 20 (95 %) estimated lower attrition rates than observed. The mean absolute difference in observed

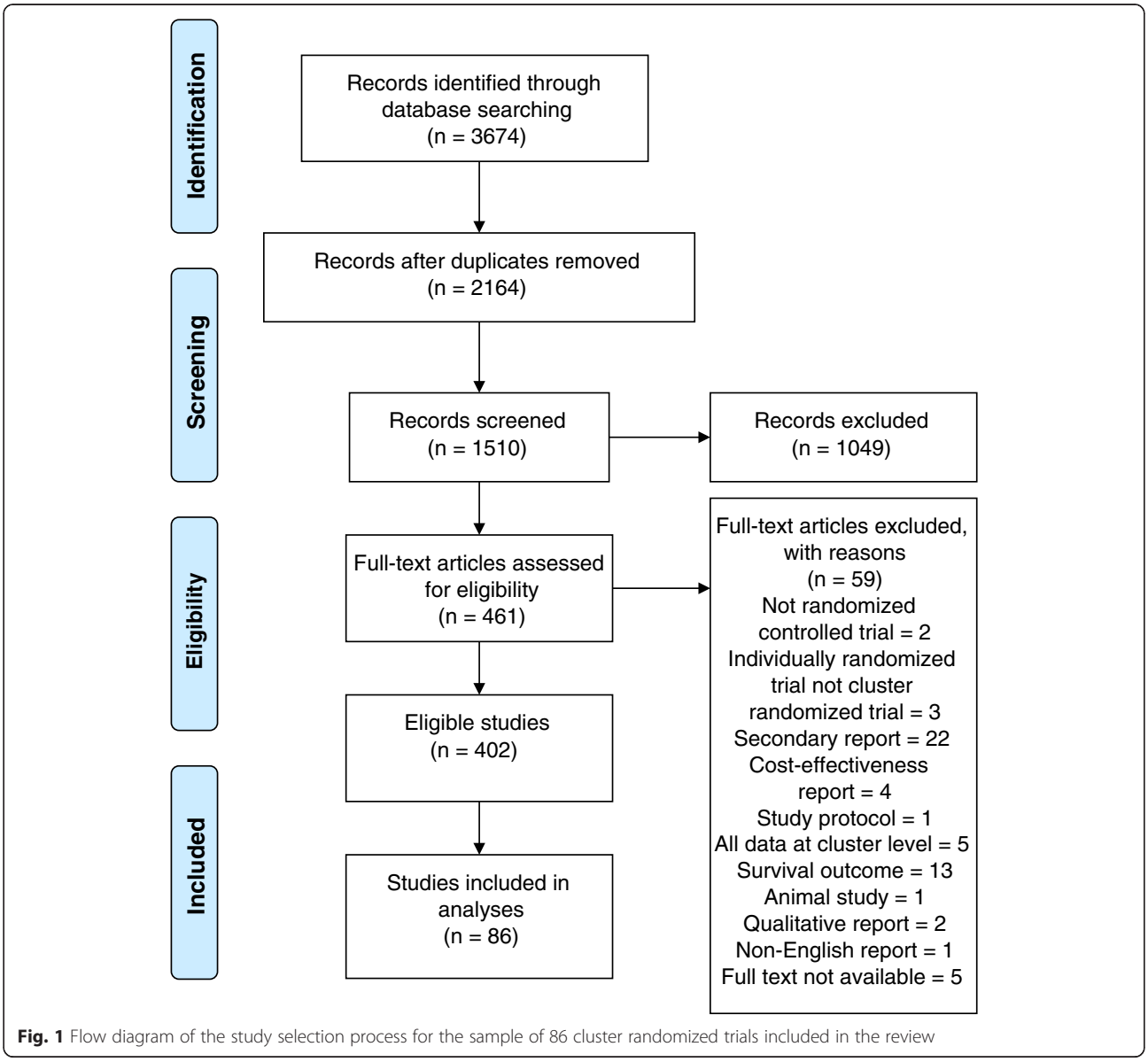

Fig. 1 Flow diagram of the study selection process for the sample of 86 cluster randomized trials included in the review

attrition rate and expected was 9 % with a range of 0.1 to 23 %.

Sensitivity analysis for missing data

Fourteen (16 %) trials reported a sensitivity analysis for missing data (Table 4), all of which reported more than 10 % missing data at the individual level. Of these, five (36 %) used MI (none of which used multilevel strategies), four (29 %) used single imputation, three (21 %) used a complete case analysis, one (7 %) used a mixed model, and one (7 %) used a mixed model with IPW.

Only five trials weakened the missingness assumption of the primary analysis to carry out their sensitivity analysis by assuming MCAR in the primary analysis and MAR in the sensitivity analysis. These five trials all used a complete case analysis as the primary analysis. For the sensitivity analysis, three of these trials used MI, one used a mixed model, and one used a mixed model with IPW. None of the trials reported using MNAR models.

Accounting for clustering in the primary analysis

The overwhelming majority of trials carried out an individual-level analysis as the primary analysis (83, 97 %). Mixed models were the most popular primary analysis used for CRTs (45, 52 %). Forty-three (96 %) of these trials accounted for clustering by adding cluster as a random effect, one trial was unclear, and one did not use cluster as a random effect. Of the 22 (26 %) trials performing an individual level basic inferential test or GLM, seven accounted for clustering via robust standard errors or design effect

Table 1 General characteristics of the 86 randomly selected cluster randomized trials published from August 2013 to July 2014

	N (%)
Stepped wedge	4 (5)
Pilot/feasibility	4 (5)
Type of outcome	
Quantitative	41 (48)
Binary	37 (43)
Count	8 (9)
How often outcome was collected	
Single	21 (24)
Repeated	65 (76)
How outcome was treated in the primary analysis	
Single	50 (58)
Repeated	36 (42)
Balance methods used in randomization	
Stratification	27 (31)[a]
Matching	14 (16)
Minimization	3 (3)
None	42 (49)
Presented sample size calculation	60 (70)

[a]One trial also used matching, and another trial also used minimization

adjustment. Fourteen (16 %) trials used GEE, with all of them accounting for clustering by using an exchangeable correlation structure. Of these, one reported estimating standard errors of parameters using the jack-knife method because the number of clusters was small [33]. One (1 %) trial carried out a descriptive analysis as the primary analysis and did not account for clustering (Table 5). Four (5 %) trials carried out a basic inferential test or GLM at the cluster level. Overall, 68 (79 %) trials accounted for clustering in the primary analysis.

Thirty-four (40 %) trials carried out an unadjusted analysis, whereas five (6 %) adjusted for balance variables only (stratification, matching, or minimization), and eight (9 %) adjusted for baseline outcome only (sometimes referred to as analysis of covariance (ANCOVA)). Thirty-nine (45 %) trials adjusted for additional covariates beyond balance variables in the primary analysis, with four of them also adjusting for baseline values of the outcome.

Table 2 Proportion of clusters with missing outcome at the primary analysis among the 86 trials included in the review

	N (%)
None	59 (69)
<10 %	14 (16)
>10 %	10 (12)
Unclear	3 (3)

Forty-six (77 %) trials reported accounting for clustering in their sample size calculations, with 41 reporting an expected ICC or CV (two trials). Of the 13 trials that reported an expected and observed ICC, seven (54 %) trials estimated larger ICCs than observed, whereas six (46 %) estimated lower ICCs than observed. The mean absolute difference in the observed and expected ICC was 0.1, with a range of 0.01 to 0.42.

Discussion

We performed a systematic review to assess how missing outcome data are being handled in CRTs. Of the 86 included CRTs, most reported some missing outcome data in the primary analysis. Among those that reported missing data, the median proportion of individuals with a missing outcome at the primary analysis was 19 %. Sixteen percent of the trials carried out a sensitivity analysis for missing data, with all of them reporting more than 10 % missing data. Only a third of these trials weakened the missingness assumption from the primary analysis.

Observed missing data rates generally exceeded expected rates, which means that researchers are not accounting enough for attrition in sample size calculations or adequately following up on participants. Furthermore, only about half (55 %) of the trials with repeated measurements used all of the outcome data in the primary analysis. Reducing repeated data to a single time point often generates a strong MCAR assumption and may reduce power. Even if the primary outcome of interest is at a particular time point, previous literature has shown that utilizing all of the information collected can minimize bias due to missing data [34].

The amount of detail in sample size calculations varied widely across trials. A few did not provide enough detail for us to indicate that a sample size calculation was performed before data collection. For example, one trial stated “sample size calculations showed 382 participants were needed.” [35] Furthermore, accounting for clustered data in sample size calculations differed among trials. One trial arbitrarily chose to increase the sample size by 30 % to account for clustering [36]. One trial stated that clustering was not accounted for in the sample size calculation because cluster sizes were expected to be small and within-cluster comparisons were not considered to be clinically meaningful [37].

Along with missing individuals, missing data can also occur at the cluster level. The removal of entire clusters with the usual solution of complete case analysis is wasteful and could lead to biased estimates depending on the missing data mechanism [38]. We did not find any studies that performed MI appropriate for clustered data (multilevel MI). Some strategies that have been

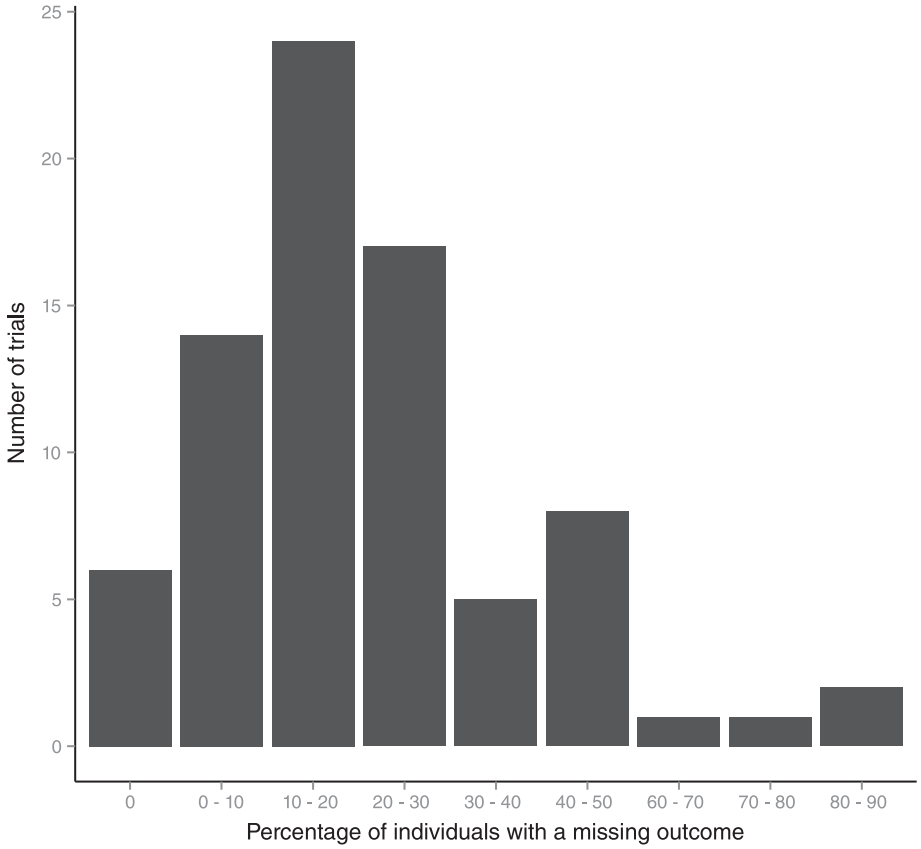

Fig. 2 Distribution of the percentage of individuals with a missing outcome for the 86 trials included in the review

Table 3 Handling of missing data in primary analysis among the 80 trials who reported missing outcome data

Methods	<10 % missing	>10 % missing	Unclear	Total
	N = 14	N = 58	N = 8	N = 80
Complete case	10	31	3	44 (55)
Single imputation				
Worst-case	1	2	0	3 (4)
LOCF	0	2	0	2 (2)
Baseline observation carried forward	0	1	0	1 (1)
Multiple imputation	0	2	0	2 (2)
GEE (unweighted)	3	0	1	4 (5)
Mixed model/ hierarchical/multilevel	0	17	1	18 (22)
Other[a]	0	0	1	1 (1)
Unclear	0	3	2	5 (6)

Abbreviations: LOCF, last observation carried forward; GEE, generalized estimating equation

[a]One trial excluded participants who dropped out or had no baseline value; for those who participated at both time points, the LOCF was carried out for a missing primary outcome

proposed to accommodate missing data in the multilevel setting, but none have been put to widespread use [15, 39–41].

In comparison to Diaz-Ordaz et al.'s [28] review, we found a higher proportion of trials reporting missing data at the cluster (28 % versus 18 %) and individual levels (93 % versus 48 %). This may be due to differences in definitions of missing data or because Diaz-Ordaz was not able to verify the amount of missing data in 31 % of the trials. We observed a similar median cluster attrition rate (7 % versus 10 %) and a slightly higher median individual attrition rate (19 % versus 13 %). Of the 95 trials with missing data, Diaz-Ordaz et al. found 66 % of the trials reporting a complete case analysis, GEE, or likelihood-based hierarchical/mixed model, whereas 18 % used single imputation and 6 % used MI. Lastly, we found a slightly higher proportion of trials reporting a sensitivity analysis for missing data (16 % versus 11 %). Compared to Bell et al.'s [7] review of 77 individually randomized controlled trials from 2013, we found a similar proportion of trials reporting missing data (93 % versus 95 %). However, CRTs were subject to higher individual level missing data rates (median 19 %, up to 90 %) compared to individually randomized trials (median 9 %, up to

Table 4 Methods for handling missing data in sensitivity analysis in 14 trials

Sensitivity method	Primary analysis	N	Total N (%)
Complete case	MI	2	3 (21)
	Mixed model	1	
Single imputation	Complete case	1	4 (29)
	Single imputation	1	
	Mixed model	2	
MI	Complete case	3	5 (36)
	Mixed model	1	
	Unclear	1	
Mixed model	Complete case	1	1 (7)
Mixed model with IPW	Complete case	1	1 (7)

Abbreviations: MI, multiple imputation; IPW, inverse probability weighting

70 %). Compared to the individually randomized trials, we found a higher proportion using complete case analysis (55 % versus 45 %) and mixed models (22 % versus 15 %). Furthermore, we found a similar proportion using GEE (4 % versus 5 %) and a lower proportion using single imputation (8 % versus 27 %) and MI (2 % versus 8 %)

More sophisticated methods are being used. Compared to a review conducted by Simpson et al. [26] of 21 CRTs from 1990 to 1993, the proportion of trials that took clustering into account in the primary analysis increased over time (57 % to 78 %). In comparison with Scott et al.'s [42] review of 150 individually randomized trials in 2001, we found a higher percentage of CRTs using stratification (31 % versus 13 %) and a similar percentage using minimization (3 % versus 4 %) compared to individually randomized trials.

Table 5 Primary analysis in 86 cluster randomized trials

	Accounted for clustering[a]		Total
Primary Analysis	Yes	No	N (%)
	N (%)	N (%)	
Individual level:			
Basic inferential test/GLM	7 (32)	15 (68)	22 (26)
GEE	14 (100)	0 (0)	14 (16)
Mixed model	43 (96)	2 (4)[b]	45 (52)
Other[c]	0 (0)	1 (100)	1 (1)
Cluster level:			
Basic inferential test/GLM	4 (100)	0 (0)	4 (5)

Abbreviations: GLM, generalized linear model; GEE, generalized estimating equation

[a]The denominator is the total number of trials performing respective primary analysis

[b]One trial was unclear

[c]Trial used a descriptive analysis as primary analysis

Our study has several strengths. Eligible studies were all CRT designs, including the stepped wedge and feasibility studies. In order to minimize the potential for bias during the review process, we had pre-specified search, study selection, and data collection strategies, all of which were carried out by two independent reviewers. We did not limit our sample space to journals with a high impact factor, thereby increasing generalizability. Three independent reviewers performed pilot testing on several trials to create a standardized data collection template. Our study has limitations as well. For example, we only chose CRTs published in English, which may result in selection bias. It was difficult to identify all CRTs because many do not include "cluster" as a term in the title or abstract. However, our search strategy included other frequently used terms for cluster randomization such as "community randomized" and "group randomized." Still, our review may have some selection bias, as researchers who do not realize their studies are cluster randomized might not follow the CONSORT guidelines, include terms such as "cluster randomized" in the title or abstract, or use robust techniques [27]. Additionally, we took a random selection of the eligible CRTs, as it was not feasible to review all 402 studies. As with any sample, this one may not be representative of the true population. However, a random selection minimizes the possibility of non-representativeness. Furthermore, we may have underestimated the amount of missing data because we used the CONSORT flow diagram, which may primarily report outcome sample size only. It is possible that missing covariates in regression models resulted in additional missing data and actual smaller sample sizes. Although some trials adjusted for additional covariates beyond balance variables, nearly all were baseline covariates such as age and gender.

In conclusion, missing data are present in the majority of CRTs, yet handling missing data in practice remains suboptimal. Appropriate methods to handle missing clustered data, particularly under the MAR assumption, should be made more accessible by methodological statisticians. For example, providing appropriate software may increase the use of such methods [43]. Moreover, researchers and applied statisticians should keep up-to-date with such methods in order to increase statistical power in trials and reduce the potential for bias. Thus, we present the following recommendations for CRTs: (1) attempt to follow up on all randomized clusters and individuals in order to limit the extent of missing data, (2) perform a primary analysis that is valid under a plausible missingness assumption and that uses all observed data, (3) perform sensitivity analyses that weaken the missing data assumption to explore the impact of departures made in the primary analysis, and (4) follow the CONSORT extension for cluster trials statement to ensure

comprehensive reporting and transparency of methods [10, 44].

Conclusions

This review aims to assess the extent and handling of missing outcome data in CRTs. Despite high rates of missing outcome data in the primary analysis, methods used to deal with missing data in practice remain inadequate. Appropriate methods, which are valid under probable missing data assumptions, should be performed to increase the statistical power and lessen the likelihood of bias. Sensitivity analysis with a weakened missing data assumption should be performed to evaluate robustness of the primary results.

Additional files

Additional file 1: PRISMA 2009 Checklist. (PDF 115 kb)

Additional file 2: References of the 86 trials included in the review. (PDF 90 kb)

Abbreviations

CI: confidence interval; CV: coefficient of variation; GEE: generalized estimating equation; GLM: generalized linear model; ICC: intracluster correlation coefficient; IPW: inverse probability weighting; MAR: missing at random; MCAR: missing completely at random; MI: multiple imputation; MNAR: missing not at random.

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

MF and MLB conceptualized the study. MF and SH collected data. MF analyzed data, drafted the manuscript, and incorporated comments from authors for successive drafts. SH, EO, and MLB contributed to the design and content. All authors read and approved the final manuscript.

Authors' information

MF is a PhD candidate in Biostatistics. SH is a PhD candidate in Biostatistics. EO is an Assistant Professor of Epidemiology. MLB is an Associate Professor of Biostatistics.

Acknowledgements

No funding was received for this study.

Received: 29 September 2015 Accepted: 28 January 2016
Published online: 09 February 2016

References

1. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold Publishers; 2000.
2. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. BMJ (Clinical research ed). 1998;317:1171–2.
3. Hayes RJ, Moulten LH. Cluster Randomised Trials. Boca Raton, FL: Chapman & Hall/CRC Press; 2009.
4. Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health. 2012;33:425–45.
5. Council NR. The Prevention and Treatment of Missing Data in Clinical Trials. Washington DC: National Academies Press. 2010.
6. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials (London, England). 2004;1:368–76.
7. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol. 2014;14:118. doi:10.1186/1471-2288-14-118.
8. Patient-Centered Outcomes Research Institute. PCORI Methodology Standards. Washington, DC: Patient Centered Outcomes Research Institute; 2012.
9. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
10. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. In: Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington D.C: National Academies Press; 2010.
11. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res. 2014;23:440–59. doi:10.1177/0962280213476378.
12. Mallinckrod CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. Drug Inform J. 2008;42:303–19.
13. Molnar FJ, Man-Son-Hing M, Hutton B, Fergusson DA. Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review Open Med. 2009;3:e31.
14. Kenward MG, Molenberghs G. Last observation carried forward: a crystal ball? J Biopharm Stat. 2009;19:872–88. doi:10.1080/10543400903105406.
15. Van Buuren S. Multiple imputation of multilevel data. In: Hox JJ, Roberts JK, editors. Handbook of advanced multilevel analysis. Milton Park: Routledge; 2011. p. 173-96.
16. Robins J, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995;90:106–21.
17. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med. 2007;26:2–19. doi:10.1002/sim.2731.
18. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846–66.
19. Cornfield J. Randomization by group: a formal analysis. Am J Epidemiol. 1978;108:100–2.
20. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Fam Pract. 2000;17:192–6.
21. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. Acad Emerg Med. 2002;9:330–41.
22. Donner A. Some aspects of the design and analysis of cluster randomization trials. J R Stat Soc: Ser C: Appl Stat. 1998;47:95–113.
23. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
24. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. New York: John Wiley & Sons; 2012.
25. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121–30.
26. Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. Am J Public Health. 1995;85:1378–83.
27. Campbell MK, Elbourne DR, Altman DG, group C. CONSORT statement: extension to cluster randomised trials. BMJ (Clinical research ed). 2004;328:702–8. doi:10.1136/bmj.328.7441.702.
28. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clin Trials (London, England). 2014. doi:10.1177/1740774514537136.
29. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62:1006–12. doi:10.1016/j.jclinepi.2009.06.005.
30. Fiero M, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomised trials: protocol for a systematic review. BMJ Open. 2015;5:e007378.
31. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28:182–91. doi:10.1016/j.cct.2006.05.007.
32. Hemming K, Haines T, Chilton P, Girling A, Lilford R. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ (Clinical research ed). 2015;350:h391.
33. Shakeshaft A, Doran C, Petrie D, Breen C, Havard A, Abudeen A, et al. The effectiveness of community action in reducing risky alcohol consumption

and harm: a cluster randomised controlled trial. PLoS Med. 2014;11: e1001617. doi:10.1371/journal.pmed.1001617.
34. Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics. An introduction to hierarchical linear modelling. Stat Med. 1999;18:855–88.
35. Freiberger E, Blank WA, Salb J, Geilhof B, Hentschke C, Landendoerfer P, et al. Effects of a complex intervention on fall risk in the general practitioner setting: a cluster randomized controlled trial. Clin Interv Aging. 2013;8:1079–88. doi:10.2147/CIA.S46218.
36. Nauta J, Knol DL, Adriaensens L, Klein Wolt K, van Mechelen W, Verhagen EA. Prevention of fall-related injuries in 7-year-old to 12-year-old children: a cluster randomised controlled trial. Br J Sports Med. 2013;47:909–13. doi:10.1136/bjsports-2012-091439.
37. Zlotkin S, Newton S, Aimone AM, Azindow I, Amenga-Etego S, Tchum K, et al. Effect of iron fortification on malaria incidence in infants and young children in Ghana: a randomized trial. JAMA. 2013;310:938–47. doi:10.1001/jama.2013.277129.
38. Campbell MJ, Walters SJ. How to design, analyse and report cluster randomised trials in medicine and health related research. Chichester: John Wiley & Sons; 2014.
39. Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. Biom J. 2008;50:329–45. doi:10.1002/bimj.200710423.
40. Ma J, Raina P, Beyene J, Thabane L. Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study. BMC Med Res Methodol. 2013;13:9. doi:10.1186/1471-2288-13-9.
41. Ma J, Akhtar-Danesh N, Dolovich L. Thabane L, investigators C. Imputation strategies for missing binary outcomes in cluster randomized trials BMC Med Res Methodol. 2011;11:18. doi:10.1186/1471-2288-11-18.
42. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. Control Clin Trials. 2002;23:662–74.
43. Pullenayegum EM, Platt RW, Barwick M, Feldman BM, Offringa M, Thabane L. Knowledge translation in biostatistics: a survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods. Stat Med. 2016;35:805–18. doi:10.1002/sim.6633. Epub 2015 Aug 25.
44. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ (Clinical research ed). 2011;342:d40.

APPENDIX C: MANUSCRIPT 3

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

Mallorie H. Fiero[1§], Chiu-Hsieh Hsu[1], Melanie L. Bell[1]

[1]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson AZ 85724

[§]Corresponding author

Contact detail:
1295 N. Martin Ave., Drachman Hall, P.O. Box 245211, Tucson, Arizona 85724
1 (520) 626-7914

Email addresses:
MHF: mfiero@email.arizona.edu
CH: pchhsu@email.arizona.edu
MLB: melaniebell@email.arizona.edu
Word count:

Summary

We extend the pattern mixture approach to handle missing 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 intracluster 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 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 process 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 are more easily understandable to 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.

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

Consider a CRT with $i = 1, ..., N$ clusters, $j = 1, ..., n_i$ individuals per cluster, and $k = 1, ..., t_{ij}$ measurements per individual. Let $Time_{ijk}$ denote the time of the kth 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}:

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

where $\gamma_i \sim N(0, \sigma^2_\gamma)$ is the random effect at the cluster level and represents deviation of each cluster from the grand mean, $\nu_{ij} \sim N(0, \mathbf{\Sigma}_\nu)$ is the random effect at the individual level and represents deviation of each individual from the cluster effect, and $\epsilon_{ijk} \sim N(0, \sigma^2_\epsilon)$ 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 between treatment arms conditionally on time.

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

$$\boldsymbol{Y_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\nu_i} + \boldsymbol{\epsilon_i}, \tag{2}$$

where $\boldsymbol{Y_i}$ 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 $N(\mathbf{0}, \mathbf{\Sigma})$ and $\boldsymbol{\epsilon_i}$ is an $N_i \times 1$ vector of random residuals distributed $N(\mathbf{0}, \sigma^2_\epsilon \boldsymbol{I_{N_i}})$, where $\boldsymbol{I_{N_i}}$ 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. Let $\mathbf{R}$ be the vector of missingness indicators for the response vector $\mathbf{Y}$ with $\mathbf{Y_{obs}}$ and $\mathbf{Y_{mis}}$ 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 process by:

$$p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \mathbf{R}|\mathbf{X}) = p(\mathbf{R}|\mathbf{X})p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{R}, \mathbf{X}), \tag{3}$$

where $p(\mathbf{R}|\mathbf{X})$ is the conditional probability distribution of the dropout pattern given observed covariates and $p(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{R}, \mathbf{X})$ 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]

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

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

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

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

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

$$\boldsymbol{\epsilon_i^*} \sim N(0, \sigma^2)$$
$$\boldsymbol{y_i^*} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\nu_i} + \boldsymbol{\epsilon_i^*}$$

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 $l = 1, ..., m$ 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 degrees of freedom $v = (m-1)(1+r^{-1})^2$. The degrees of freedom depends on m and the ratio $r = (1+m^{-1})B/\bar{W}$, which is the relative increase in variance due to missing data [17].

3 Simulation study

3.1 Data generation

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 [23]:

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

where $\gamma_i \sim N(0, \sigma_\gamma^2)$ denotes the random cluster effect, $\nu_{ij} \sim N(0, \mathbf{\Sigma}_\nu)$ 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. 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 in pairs as (12, 30), (12, 100), (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 a true $k = 1.0$ for the control arm and $k = 1.75$ for the treatment arm, 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.

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 [24]. 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 $k = 0.8$ dereases the imputed values by 20%, and incorrectly assumes that the unobserved outcomes are lower (i.e., better outcome) than the observed outcomes [23].

Using the completed dataset, we modeled the outcome with a mixed model (`lme4`) 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,

$$([0.60 \times (\beta_0 + \beta_1 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_3)]) - \beta_0 = -1.8$$

and the true treatment effect was,

$$([0.60 \times (\beta_0 + \beta_1 + \beta_3)] + [0.40 \times (\beta_0 + \beta_1 + \beta_3)]) - (\beta_0 + \beta_1) = -0.8.$$

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. Empirical standard error: standard deviation of mean across samples
4. Ratio of model-based to empirical standard error

3.3 Results

We present the results of our simulations in Tables 1 - 4. 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 -3.02% and -9.34%, 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 -149.67% and -189.77%, respectively. Percent bias is more extreme in the treatment effect. For example, the percent bias for $k = 1.3$ is -38.66% for change over time in the treatment arm, and -89.53% 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.

Tables 3 - 4 display the empirical standard errors and ratios of model-based to empirical standard errors for change over time in the treatment arm and treatment effect. Overall, results were similar for the percent bias of the treatment arm change over time and treatment effect. Larger 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 $k = 1.7$.

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 [25]. 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 5 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 completers and non-completers 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 ($y_{2_T} - y_{2_C}$). Inferences were combined using Rubin's rules.

4.3 Results

Table 6 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 is prevalent in CRTs. It is crucial to accommodate for 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 who indicated using 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 [26].

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 [27]. 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 [28]. A range of plausible k or δ can be specified, or an average can be specified if a single analysis is preferred.

Multiplying imputed MAR values by k to create MNAR values can be implemented in both treatment arms or in one treatment arm only. 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. Extending the PMM to a CRT with more than two time points becomes more challenging. The multiplier k can be specified at each time point 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 [29, 30]. 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. 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.

24. Van Buuren S, Oudshoorn C. mice: Multivariate imputation by chained equations. r package version 1.16 2007.

25. 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.

26. Mistler SA. A sas macro for applying multiple imputation to multilevel data. *Proceedings of the SAS Global Forum*, 2013.

27. 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.

28. 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.

29. Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* 1991; **10**(4):585–598.

30. 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	30	0.001	-82.77	-64.94	-38.20	-2.55	-186.79	-146.68	-86.52	-6.31
		0.01	-83.21	-65.39	-38.66	-3.02	-189.77	-149.67	-89.53	-9.34
		0.1	-85.04	-67.47	-41.13	-6.01	-195.05	-155.54	-96.26	-17.24
		0.3	-84.75	-66.68	-39.59	-3.46	-180.70	-140.06	-79.09	2.21
		0.5	-84.37	-66.75	-40.32	-5.07	-182.25	-142.61	-83.14	-3.84
12	100	0.001	-83.75	-65.91	-39.14	-3.45	-186.73	-146.58	-86.36	-6.06
		0.01	-83.51	-65.55	-38.61	-2.69	-183.83	-143.42	-82.80	-1.98
		0.1	-84.53	-66.73	-40.05	-4.47	-185.90	-145.87	-85.82	-5.76
		0.3	-84.15	-66.48	-39.98	-4.64	-195.72	-155.97	-96.34	-16.83
		0.5	-85.24	-67.26	-40.28	-4.32	-189.63	-149.17	-88.48	-7.56
30	30	0.001	-83.96	-66.08	-39.26	-3.51	-188.81	-148.58	-88.24	-7.78
		0.01	-85.99	-68.29	-41.74	-6.33	-190.47	-150.63	-90.88	-11.22
		0.1	-84.41	-66.66	-40.05	-4.56	-192.69	-152.76	-92.87	-13.02
		0.3	-83.97	-66.55	-40.42	-5.57	-198.89	-159.69	-100.89	-22.49
		0.5	-84.74	-66.89	-40.13	-4.44	-204.15	-163.99	-103.77	-23.46
30	100	0.001	-83.93	-66.20	-39.61	-4.15	-190.23	-150.33	-90.50	-10.71
		0.01	-84.35	-66.61	-40.01	-4.54	-190.29	-150.38	-90.53	-10.72
		0.1	-84.87	-66.98	-40.14	-4.36	-184.42	-144.16	-83.78	-3.27
		0.3	-84.28	-66.29	-39.31	-3.34	-183.15	-142.69	-81.99	-1.05
		0.5	-83.57	-65.76	-39.03	-3.39	-190.36	-150.27	-90.14	-9.95

Abbreviations: MNAR, missing not at random; ICC, intracluster correlation coefficient
[1]Treatment 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	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	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, intracluster correlation coefficient

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

Table 3: Empirical standard errors for change over time in the treatment arm and treatment effect.

No. clusters	Cluster size	ICC	Treatment arm change over time	Treatment effect[1]
12	30	0.001	0.224	0.335
		0.01	0.234	0.371
		0.1	0.234	0.666
		0.3	0.231	1.147
		0.5	0.234	1.848
12	100	0.001	0.128	0.182
		0.01	0.122	0.245
		0.1	0.129	0.609
		0.3	0.125	1.214
		0.5	0.126	1.781
30	30	0.001	0.364	0.508
		0.01	0.372	0.592
		0.1	0.365	1.078
		0.3	0.366	1.984
		0.5	0.386	2.847
30	100	0.001	0.212	0.307
		0.01	0.204	0.416
		0.1	0.203	0.936
		0.3	0.201	1.778
		0.5	0.190	2.757

Abbreviations: ICC, intracluster correlation coefficient

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

Table 4: 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	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	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, intracluster correlation coefficient

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

Table 5: PoNDER study. Means and standard deviations of baseline EPDS score by treatment arm and dropout pattern.

	Control N = 1151		Treatment N = 2268	
Dropout pattern	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 6: 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
[1]Treatment effect: mean difference between treatment arms at follow-up

APPENDIX D: MANUSCRIPT 4

Comparison of missing data strategies for missing cluster level covariates: a simulation study

Mallorie H. Fiero[1§], Melanie L. Bell[1]

[1]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson AZ 85724

[§]Corresponding author

Contact detail:
1295 N. Martin Ave., Drachman Hall, P.O. Box 245211, Tucson, Arizona 85724
1 (520) 626-7914

Email addresses:
MHF: mfiero@email.arizona.edu
MLB: melaniebell@email.arizona.edu
Word count:

Summary

In multilevel data, subjects are grouped within clusters, such as students within schools or patients within medical practices. A high proportion of missing data can occur among multilevel data because missing data can be found at the individual level as well as the cluster level. Missing data among cluster level covariates has received little attention in the literature. Therefore, we performed a simulation study to evaluate the performance of missing data strategies for missing cluster level covariates under the missing at random assumption. We consider (1) the linear mixed effects model, (2) single imputation: mean substitution (continuous variable) or mode substitution (categorical variable), (3) single level multiple imputation (MI) ignoring clustering, (4) MI incorporating clusters as fixed effects, and (5) MI at the cluster level using aggregated data. Our results show that when the intracluster correlation coefficient (ICC) is small ($\rho \leq 0.1$) and the percentage of missing cluster level data is low ($\leq 25\%$), the mixed model produces unbiased estimates of the regression coefficients and ICC. Otherwise, MI at the cluster level using aggregated

data performs well when handling missing cluster level data, though caution should be taken if the percentage of missing data is high.

KEY WORDS: missing data; multilevel data; multiple imputation

1 Introduction

In multilevel data, subjects are grouped within clusters. Some examples of clusters include schools, medical practices, and communities. Subjects within a cluster are likely to be more similar, leading to decreased statistical power compared to independent data. The intracluster correlation coefficient (ICC) measures the similarity of subjects within clusters and is defined as the proportion of variance due to clustering. Failing to account for clustering in statistical analysis leads to an underestimation of standard errors and artificially narrow confidence intervals [1, 2].

Empirically, the ICC is very low for clustered data where subjects are grouped within clusters. For example, Bell and McKenzie [3] assessed 87 ICCs from 15 psycho-oncology cluster randomized trials, which randomize clusters to treatment arms rather than individuals, and found the median ICC to be 0.02. Adams et al. [4] examined 1039 ICCs from 31 multilevel studies found in primary care research, and reported the median ICC to be 0.01. Along with subjects grouped within clusters, multilevel data also occur in longitudinal data, where individuals are measured repeatedly over time. ICCs found within longitudinal data are generally higher. We focus our attention on non-longitudinal multilevel data where subjects are grouped within clusters and the ICC is lower.

Missing data is a common problem in multilevel studies and should be accommodated with appropriate statistical techniques, as it leads to decreased statistical power and bias in some cases [5]. Missing data mechanisms are crucial when choosing an appropriate approach to handle missing data. They have been categorized into three groups, which describe the relationships between the probability of missingness and the observed and unobserved data. Data are missing completely at random (MCAR) when the probability of missingness does not depend on the observed outcomes and covariates. Data are considered missing at random (MAR) when missingness is independent of the unobserved data after conditioning on observed data. Lastly, data are missing not at random (MNAR) when the probability of missingness depends on the missing value even after conditioning on the observed data [6].

The proportion of missing data can be higher in multilevel data compared to independent data because missing data can occur at the individual level and cluster level. In an educational dataset where students are grouped by classroom, missing data can occur at the student (or individual) level, such as age of the student or test score. Missing data can

also occur at the classroom (or cluster) level, such as teacher's highest level of education. Missing individual level outcomes and covariates have been considered by many. Some references include [7, 8, 9, 10, 11]. However, missing data among cluster level covariates have received limited attention. Gibson et al. [12] and Cheung [13] both evaluated strategies to handle continuous missing cluster level covariates under MCAR, such as complete case analysis, single level multiple imputation (MI) ignoring clustering, and mean substitution. Both studies found complete case analysis to perform well with less than 50% missing cluster level data. A complete case analysis, also known as listwise deletion, eliminates observations with missing data. It can be very inefficient when missing cluster level covariates are present, since all of the observations within the cluster are removed. Additionally, this method loses precision since information is being deleted, and can produce biased estimates if the missing data mechanism is not MCAR. Both studies found single level MI ignoring clustering to be a poor strategy when implemented within the multilevel structure, because it produces underestimated standard errors.

When imputing cluster level variables, imputed values must be the same within each cluster. Gelman and Hill [14] suggested an approach using MI to impute missing cluster level covariates, which involves separating the data into subject level and cluster level datasets and imputing within each dataset. The imputed data are then combined to create complete datasets and analyzed for inference. This method, however, has not yet been compared to other commonly used techniques to impute missing cluster level data, such as the linear mixed effects model, mean substitution, and other MI approaches. We extend the investigation of missing cluster level covariates by performing a simulation study to assess the performance of missing data strategies under MAR, as this assumption may be more reasonable in practice. We examine the sensitivity of methods to the total number of clusters, cluster size (number of subjects per cluster), ICC, and percentage of missingness. Since cluster level covariates are often found to be categorical, we evaluate these methods when the missing cluster level covariate is categorical as well as continuous.

2 Methods

We used a simulation study designed after Van Buuren's previous work in missing multilevel data [9] to investigate the performance of strategies to handle missing cluster level covariates under the MAR assumption. This section contains an overview of commonly used methods to accommodate missing continuous and categorical cluster level data, followed by an introduction to the linear mixed effects model, which we use to compute regression coefficients and variance components after performing each missing data method.

2.1 Missing data methods

There are several methods that can be used to handle missing cluster level covariates, including the linear mixed effects model, mean substitution (continuous variable), mode substitution (categorical variable), single level MI ignoring clustering, fixed effects MI, and MI aggregate imputation. It is unclear which methods are being used in practice to handle missing cluster level covariates.

The linear mixed effects model (mixed model) uses a likelihood-based approach to estimate parameters. Under the correct model specification, the mixed model performs an implicit imputation and produces unbiased estimates under MAR if outcome data are missing. This method excludes observations with missing cluster level covariates from the analysis, and can produce biased estimates in the presence of missing covariate variables even under MCAR [15].

Among continuous cluster level covariates, mean substitution replaces missing observations with the overall mean across observed clusters. For categorical variables, mode substitution replaces missing data with the most common category found among the observed data. Similar to other single imputation techniques, which replace missing observations with a single value, these methods are prone to underestimation of variance. Although simple, these methods do not condition on any other information in the observation, which can generate misleading relationships between variables [9].

The MI procedure can be separated into two steps: (1) imputation of missing data and (2) analysis of complete multiply imputed datasets. We carry out MI using the `mice` package in R, which uses the multivariate imputation by chained equations (MICE) technique [16]. Imputation models are used for each variable with missing data so that unobserved values are imputed based on a conditional distribution of other variables with observed data to create m complete datasets. For a continuous variable with missing data, the imputation model is a linear regression model fitted using observed data from covariates. The discriminant function method is used to impute a factor variable with more than two categories, assuming the covariates in the imputation model are approximately multivariate normal and covariance matrices are equal across groups [16, 17]. After the imputation step, each completed dataset is then analyzed using standard statistical techniques and combined for inference [18]. Another popular MI approach is the Markov chain Monte Carlo (MCMC) method, which draws imputations assuming the missing data follow a multivariate normal distribution. This method is appealing when a multivariate normal distribution can be specified for the data. On the other hand, MICE may be more suitable when a multivariate distribution cannot be specified to describe the data.

Standard MI, which we will call single level MI, assumes the observations are independent, and ignores the clustered structure found in multilevel data. This results in underestimated standard errors and confidence intervals that are too narrow [9]. One approach to incor-

porate clustering in the MI procedure, that we will call fixed effects MI, includes cluster as a fixed effect in the imputation model when performing MI, which models the differences in intercepts between the clusters [9].

However, neither single level MI nor fixed effects MI will impute the same value for all individuals within a single cluster, which is inappropriate when imputing missing cluster level variables. Gelman and Hill [14] proposed using a MI aggregate imputation approach, which imputes the same value for each individual within a cluster. This method involves separating the data into two datasets: one individual level dataset and one cluster level dataset. In order to impute missing cluster level data, the individual level data are first aggregated into cluster level summaries (such as the cluster mean). The aggregated individual level data are then incorporated with the cluster level data for imputation, so that the combined dataset includes a single row for each cluster with the cluster level variables and the aggregated individual level variables. The aggregated individual level variables are included in the imputation model to impute a single value for each missing cluster level covariate. The imputed cluster level covariates are then combined with the original individual level data, so that the final dataset includes the cluster level variables with the same imputed value across individuals within a cluster. Each completed multiply imputed dataset is then analyzed and combined for inference [14].

2.2 Linear mixed effects model

The linear mixed effects model is often used to analyze clustered data. We use this method to analyze the completed dataset after performing each imputation approach. Consider a multilevel dataset with $i = 1, ..., N$ clusters and $j = 1, ..., n_i$ subjects per cluster. A single outcome of interest Y_{ij} is modeled by:

$$Y_{ij} = X_{ij}\beta + U_i + e_{ij} \tag{1}$$

where Y_{ij} is an $n_i \times 1$ vector of responses, X_{ij} is a known $n_i \times p$ design matrix of fixed effects at the individual level or cluster level, β is a $p \times 1$ vector of unknown fixed effects, U_i represents the random cluster effects distributed $N(0, \sigma_B^2)$, and e_{ij} represents the individual error terms distributed $N(0, \sigma_W^2)$. Additionally, U_i and e_{ij} are assumed to be uncorrelated. The variance of Y_{ij} is $\sigma^2 = \sigma_B^2 + \sigma_W^2$, where σ_B^2 denotes the between cluster variance and σ_W^2 denotes the within cluster variance. The ICC $= \sigma_B^2/\sigma^2$.

3 Simulation study

In this section, we describe our simulation study to assess methods to handle missing continuous and categorical cluster level covariates.

3.1 Continuous cluster level covariate

Multilevel data with a continuous cluster level covariate, W_i, were generated using the following model [9]:

$$Y_{ij} = \beta_0 + \beta_1 W_i + U_i + e_{ij} \tag{2}$$

with $U_i \sim N(0, \sigma^2_W)$ and $e_{ij} \sim N(0, \sigma^2_B)$ denoting the random cluster effects and measurement error terms, respectively. The regression coefficients were defined as $\beta_0 = 0$ and $\beta_1 = 0.5$. We set the variance parameters $\sigma^2 = \sigma^2_W + \sigma^2_B = 0.75$, and varied the ICC = (0.001, 0.01, 0.1, 0.3). We simulated both small and large sample sizes by varying the total number of clusters $N = (24, 60)$ and cluster size $n_i = (20, 50)$.

The cluster level covariate W_i was deleted under the MAR assumption with 25% and 50% missing. For the cluster average Y_i less than than the upper 25th (or 50th) percentile of the standard normal distribution, the non-response probability in W_i was 10%. For the cluster average Y_i greater than or equal to the upper 25th (or 50th) percentile of the standard normal distribution, the non-response probability in W_i was 90%.

We simulated 1000 replications from each scenario described above, and performed the following methods to handle missing cluster level continuous covariates: the mixed model (MM), mean substitution (MN), single level MI ignoring clustering (SL), fixed effects MI (FE), MI aggregate imputation (AG). All MI methods were carried out as described in the previous section using the `mice` package in `R` version 3.2.3 [19]. Imputations were fixed to 20 with the outcome variable included in the imputation model. For each completed dataset, we modeled the outcome with a mixed model via the R package `lme4` using Equation 2. Fixed and random parameter estimates were pooled using Rubin's Rules [18] as implemented in `mice`. For each missing data method, we calculated the estimated fixed and random effects parameters and examined their bias, defined as the difference between the estimate of the parameter and the true value. Coverage was also computed as the proportion of times the 95% confidence interval contained the true value of the fixed parameter estimate.

3.2 Categorical cluster level covariate

When the cluster level covariate was categorical, we simulated the data similarly, except with the following multilevel model:

$$Y_{ij} = \beta_0 + \beta_1(\text{Group 2}) + \beta_2(\text{Group 3}) + U_i + e_{ij}$$

where the regression coefficients were defined as $\beta_0 = 0$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. With β_1 and β_2 set as dummy variables, this creates three cluster level groups with means 0 (group

1), 0.5 (group 2), and 0.9 (group 3). There was an equal probability of being assigned to one of the three groups. The following methods used to handle missing cluster level categorical data were examined: the mixed model (MM), mode substitution (MD), single level MI ignoring clustering (SL), fixed effects MI (FE), MI aggregate imputation (AG). We deleted the categorical cluster level covariate under MAR and assessed performance of methods similar to the continuous cluster level covariate case.

4 Results

4.1 Continuous cluster level covariate

The results of our simulations for missing data in the continuous cluster level covariate are presented in Tables 1 - 3. When the ICC was small (ICC ≤ 0.1) with 25% missing data, the mixed model performed best, as it yielded unbiased regression coefficients and reasonable coverage rates. For example, the estimated β_0 was -0.01 and β_1 was 0.49 for 24 clusters, 50 subjects per cluster, an ICC of 0.01, and 25% missing data (Table 1). Compared to the other missing data strategies, the mixed model generated the closest ICC estimates to the true value, particularly when the ICC was lower. However, when the ICC was higher (ICC > 0.1), the mixed model produced severely underestimated regression coefficients and performed worse with an increased amount of missing data. For 60 clusters, 20 subjects per cluster, an ICC of 0.3, and 50% missing data, the estimated β_0 was -0.33 and β_1 was 0.36 (Table 1).

MI aggregate imputation performed best when the ICC was larger. It generated consistently close fixed parameter estimates to the true values as well as adequate coverage of the fixed parameters. For example, coverage for β_0 and β_1 was 94% when there were 24 clusters, 20 subjects per cluster, an ICC of 0.1, and 25% missing data (Table 3). Similar to the other imputation approaches, MI aggregate imputation tended to overestimate the ICC particularly with 50% missingness, though it generally performed better. When there were 60 clusters, 50 subjects per cluster, an ICC of 0.3, and 25% missing, MI aggregate imputation estimated the ICC to be 0.36, while mean substitution, single level MI, and fixed effects MI estimated the ICC to be 0.46, 0.42, and 0.42, respectively (Table 2).

The worst method was mean substitution, which became more apparent with more missing cluster level covariates. Mean substitution severely overestimated the β_0 coefficient as well as the ICC. For example, the estimated β_0 was 0.29 and ICC was 0.19 for 24 clusters, 50 subjects per cluster, an ICC of 0.001, and 50% missing data (Tables 1 - 2). Single level MI and fixed effects MI performed better than mean substitution, but were under-coveraged for the fixed effects due to underestimation of the standard errors. The results did not change substantially when varying the total number of clusters or cluster size.

Table 1: Estimates of the regression coefficients based on methods to handle 25% and 50% missing data in continuous cluster level covariates

No. clusters	Cluster size	ICC	$\beta_0 = 0$ MM	MN	SL	FE	AG	$\beta_1 = 0.5$ MM	MN	SL	FE	AG
25% missing												
24	20	0.001	-0.01	0.08	0.07	0.07	0.05	0.48	0.48	0.48	0.48	0.53
24	20	0.01	-0.02	0.08	0.07	0.07	0.05	0.47	0.47	0.47	0.47	0.53
24	20	0.1	-0.05	0.08	0.06	0.06	0.05	0.43	0.43	0.39	0.39	0.49
24	20	0.3	-0.14	0.07	0.03	0.03	0.04	0.38	0.38	0.27	0.27	0.43
24	50	0.001	-0.01	0.08	0.06	0.06	0.04	0.49	0.49	0.47	0.47	0.54
24	50	0.01	-0.01	0.08	0.06	0.06	0.04	0.49	0.49	0.46	0.46	0.54
24	50	0.1	-0.05	0.08	0.05	0.05	0.04	0.44	0.44	0.38	0.38	0.5
24	50	0.3	-0.14	0.07	0.02	0.02	0.03	0.40	0.40	0.25	0.25	0.46
60	20	0.001	-0.02	0.08	0.06	0.06	0.03	0.48	0.48	0.48	0.48	0.53
60	20	0.01	-0.02	0.08	0.06	0.06	0.03	0.47	0.47	0.47	0.47	0.53
60	20	0.1	-0.06	0.08	0.05	0.05	0.03	0.43	0.43	0.40	0.40	0.50
60	20	0.3	-0.14	0.08	0.03	0.03	0.03	0.39	0.39	0.26	0.26	0.47
60	50	0.001	-0.01	0.08	0.06	0.06	0.03	0.49	0.49	0.47	0.47	0.53
60	50	0.01	-0.01	0.08	0.06	0.06	0.03	0.49	0.49	0.46	0.46	0.53
60	50	0.1	-0.05	0.08	0.05	0.05	0.03	0.44	0.44	0.37	0.37	0.51
60	50	0.3	-0.14	0.08	0.03	0.03	0.03	0.39	0.39	0.24	0.24	0.47
50% missing												
24	20	0.001	-0.08	0.26	0.20	0.20	0.20	0.45	0.45	0.41	0.41	0.51
24	20	0.01	-0.09	0.26	0.20	0.20	0.19	0.44	0.44	0.41	0.41	0.49
24	20	0.1	-0.19	0.20	0.13	0.13	0.13	0.39	0.39	0.33	0.33	0.41
24	20	0.3	-0.34	0.15	0.05	0.05	0.08	0.35	0.35	0.24	0.24	0.36
24	50	0.001	-0.04	0.29	0.22	0.22	0.23	0.47	0.47	0.42	0.42	0.54
24	50	0.01	-0.06	0.28	0.22	0.22	0.22	0.46	0.46	0.42	0.42	0.53
24	50	0.1	-0.18	0.20	0.13	0.13	0.14	0.39	0.39	0.32	0.32	0.42
24	50	0.3	-0.33	0.15	0.06	0.06	0.08	0.35	0.35	0.23	0.23	0.35
60	20	0.001	-0.07	0.27	0.20	0.20	0.13	0.45	0.45	0.42	0.42	0.53
60	20	0.01	-0.09	0.26	0.20	0.20	0.13	0.45	0.45	0.41	0.41	0.52
60	20	0.1	-0.19	0.21	0.13	0.13	0.10	0.40	0.40	0.35	0.34	0.47
60	20	0.3	-0.33	0.15	0.06	0.06	0.06	0.36	0.36	0.24	0.24	0.42
60	50	0.001	-0.03	0.29	0.22	0.22	0.14	0.48	0.48	0.43	0.43	0.56
60	50	0.01	-0.05	0.28	0.21	0.21	0.14	0.47	0.47	0.41	0.41	0.54
60	50	0.1	-0.17	0.21	0.13	0.13	0.10	0.41	0.41	0.33	0.33	0.48
60	50	0.3	-0.33	0.15	0.06	0.06	0.06	0.36	0.36	0.23	0.23	0.42

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 2: Estimates of the intracluster correlation coefficient and residual variance based on methods to handle 25% and 50% missing data in continuous cluster level covariate

No. clusters	Cluster size	ICC	MM	MN	SL	FE	AG	σ^2	MM	MN	SL	FE	AG
25% missing													
24	20	0.001	0.00	0.11	0.08	0.08	0.04	0.749	0.74	0.75	0.73	0.73	0.74
24	20	0.01	0.01	0.12	0.09	0.09	0.05	0.74	0.74	0.74	0.72	0.72	0.74
24	20	0.1	0.08	0.24	0.19	0.19	0.16	0.65	0.65	0.65	0.64	0.64	0.65
24	20	0.3	0.23	0.47	0.41	0.41	0.39	0.45	0.45	0.45	0.44	0.44	0.45
24	50	0.001	0.00	0.11	0.08	0.08	0.03	0.749	0.75	0.75	0.73	0.73	0.75
24	50	0.01	0.01	0.11	0.08	0.08	0.04	0.74	0.74	0.74	0.72	0.72	0.74
24	50	0.1	0.09	0.23	0.19	0.19	0.15	0.65	0.65	0.65	0.64	0.64	0.65
24	50	0.3	0.24	0.47	0.43	0.43	0.39	0.45	0.45	0.45	0.44	0.44	0.45
60	20	0.001	0.00	0.11	0.07	0.07	0.02	0.749	0.75	0.75	0.73	0.73	0.75
60	20	0.01	0.01	0.12	0.08	0.08	0.03	0.74	0.74	0.74	0.72	0.72	0.74
60	20	0.1	0.08	0.23	0.19	0.19	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	20	0.3	0.23	0.46	0.41	0.41	0.36	0.45	0.45	0.45	0.44	0.44	0.45
60	50	0.001	0.00	0.10	0.07	0.07	0.02	0.749	0.75	0.75	0.73	0.73	0.75
60	50	0.01	0.01	0.11	0.08	0.08	0.03	0.74	0.74	0.74	0.72	0.72	0.74
60	50	0.1	0.09	0.23	0.19	0.19	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	50	0.3	0.23	0.46	0.42	0.42	0.36	0.45	0.45	0.45	0.44	0.44	0.45
50% missing													
24	20	0.001	0.01	0.19	0.16	0.16	0.12	0.749	0.74	0.75	0.71	0.71	0.75
24	20	0.01	0.01	0.21	0.17	0.17	0.14	0.74	0.73	0.74	0.70	0.70	0.74
24	20	0.1	0.07	0.30	0.25	0.25	0.24	0.65	0.65	0.65	0.62	0.62	0.65
24	20	0.3	0.20	0.52	0.45	0.45	0.45	0.45	0.45	0.45	0.43	0.43	0.45
24	50	0.001	0.00	0.19	0.15	0.15	0.12	0.749	0.75	0.75	0.71	0.71	0.75
24	50	0.01	0.01	0.20	0.17	0.17	0.14	0.74	0.74	0.74	0.70	0.70	0.74
24	50	0.1	0.08	0.30	0.25	0.25	0.24	0.65	0.65	0.65	0.62	0.62	0.65
24	50	0.3	0.21	0.51	0.45	0.45	0.46	0.45	0.45	0.45	0.43	0.43	0.45
60	20	0.001	0.00	0.19	0.14	0.14	0.08	0.749	0.74	0.75	0.70	0.70	0.75
60	20	0.01	0.01	0.20	0.15	0.15	0.09	0.74	0.74	0.74	0.69	0.69	0.74
60	20	0.1	0.07	0.29	0.24	0.24	0.19	0.65	0.65	0.65	0.61	0.61	0.65
60	20	0.3	0.21	0.50	0.43	0.43	0.40	0.45	0.45	0.45	0.43	0.43	0.45
60	50	0.001	0.00	0.18	0.14	0.14	0.07	0.749	0.75	0.75	0.70	0.70	0.75
60	50	0.01	0.01	0.19	0.15	0.15	0.08	0.74	0.74	0.74	0.69	0.69	0.74
60	50	0.1	0.08	0.29	0.24	0.24	0.19	0.65	0.65	0.65	0.61	0.61	0.65
60	50	0.3	0.21	0.50	0.43	0.43	0.40	0.45	0.45	0.45	0.43	0.43	0.45

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 3: Coverage of true values by the 95% confidence interval of the regression coefficients based on methods to handle 25% and 50% missing data in continuous cluster level covariate

No. clusters	Cluster size	ICC	β_0 MM	MN	SL	FE	AG	β_1 MM	MN	SL	FE	AG
25% missing												
24	20	0.001	95	88	87	87	89	94	100	98	98	96
24	20	0.01	94	87	87	87	90	91	99	96	96	95
24	20	0.1	88	92	92	92	94	86	96	82	80	94
24	20	0.3	79	93	93	93	93	82	93	47	47	94
24	50	0.001	94	87	87	87	89	95	100	97	97	94
24	50	0.01	93	89	90	90	90	92	100	95	96	95
24	50	0.1	89	91	92	92	93	87	97	63	63	97
24	50	0.3	77	93	92	92	95	85	96	20	20	95
60	20	0.001	94	67	72	72	87	90	100	98	98	90
60	20	0.01	90	69	74	74	87	88	100	96	96	92
60	20	0.1	81	81	88	88	90	76	94	61	61	94
60	20	0.3	56	86	92	92	93	73	90	7	8	94
60	50	0.001	94	60	64	64	79	93	100	95	95	79
60	50	0.01	92	66	72	72	83	91	100	89	90	82
60	50	0.1	80	78	87	87	90	81	96	21	22	94
60	50	0.3	55	89	94	94	92	71	90	1	1	94
50% missing												
24	20	0.001	84	68	59	59	64	92	100	89	91	97
24	20	0.01	79	69	59	59	69	89	100	87	86	97
24	20	0.1	54	82	80	80	88	75	98	67	66	94
24	20	0.3	39	92	91	91	94	74	96	42	43	93
24	50	0.001	90	55	36	35	47	93	100	80	79	97
24	50	0.01	81	60	44	44	55	89	100	75	75	97
24	50	0.1	55	82	76	76	88	78	99	47	48	96
24	50	0.3	39	90	89	89	92	76	95	29	30	92
60	20	0.001	64	9	9	8	41	83	100	77	77	91
60	20	0.01	54	13	14	14	48	78	100	70	71	93
60	20	0.1	28	46	56	57	75	66	95	41	40	92
60	20	0.3	12	77	88	89	91	63	92	13	13	89
60	50	0.001	83	2	1	2	28	91	100	60	61	83
60	50	0.01	72	4	3	3	35	82	100	46	46	90
60	50	0.1	30	42	51	50	70	67	97	15	15	92
60	50	0.3	12	76	88	88	92	63	89	7	7	87

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MN = mean substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

4.2 Categorical cluster level covariate

The results of our simulations for missing data in categorical cluster level covariates are presented in Tables 4 - 6. With 25% missing data and lower ICC (ICC ≤ 0.1), the best methods were the mixed model and MI aggregate imputation. Both strategies generated reasonable fixed parameter estimates, coverage, and ICC estimates. For 60 clusters, 20 subjects per cluster, an ICC of 0.001, and 25% missing data, the mixed model estimated β_0 as 0, β_1 as 0.50, and β_2 as 0.89, while MI aggregate imputation estimated β_0 as 0, β_1 as 0.50, and β_2 as 0.90 (Table 4). The mixed model had problems with convergence when the total number of clusters was lower ($N = 24$). When the ICC was higher, MI aggregate imputation outperformed all other strategies. However, when the amount of missing data increased to 50%, none of the missing data strategies performed particularly well. Overall, the worst method to handle missing categorical cluster level covariates was mode substitution. The fixed parameter estimates were extremely biased, especially with higher ICC. Mode substitution also overestimated the ICC, and became worse with 50% missing data. For 24 clusters, 20 subjects per cluster, an ICC of 0.01, and 25% missing data, mode substitution estimated the ICC to be 0.06, which increased to 0.14 when the percentage of missing data increased to 50% (Table 5).

Table 4: Estimates of the regression coefficients based on methods to handle 25% and 50% missing data in categorical cluster level covariates

No. clusters	Cluster size	ICC	$\beta_0 = 0$					$\beta_1 = 0.5$					$\beta_2 = 0.9$				
			MM	MD	SL	FE	AG	MM	MD	SL	FE	AG	MM	MD	SL	FE	AG
25% missing																	
24	20	0.001	0.00	0.06	0.02	0.03	0.00	0.50	0.45	0.50	0.49	0.51	0.87	0.79	0.87	0.86	0.89
24	20	0.01	0.00	0.06	0.02	0.03	0.00	0.50	0.45	0.50	0.50	0.51	0.86	0.78	0.85	0.85	0.89
24	20	0.1	-0.01	0.11	0.09	0.09	0.01	0.48	0.41	0.46	0.45	0.52	0.77	0.66	0.73	0.74	0.84
24	20	0.3	-0.03	0.17	0.21	0.20	0.07	0.41	0.31	0.31	0.30	0.46	0.66	0.49	0.48	0.50	0.73
24	50	0.001	0.00	0.04	0.01	0.02	0.00	0.50	0.46	0.49	0.49	0.50	0.90	0.83	0.88	0.88	0.90
24	50	0.01	0.00	0.05	0.02	0.03	0.00	0.50	0.46	0.49	0.48	0.50	0.88	0.81	0.86	0.86	0.89
24	50	0.1	0.00	0.11	0.11	0.11	0.02	0.47	0.40	0.43	0.42	0.51	0.78	0.67	0.69	0.69	0.84
24	50	0.3	-0.02	0.16	0.23	0.23	0.06	0.41	0.32	0.28	0.28	0.46	0.67	0.50	0.43	0.46	0.73
60	20	0.001	0.00	0.07	0.02	0.02	0.00	0.50	0.45	0.50	0.50	0.50	0.86	0.78	0.87	0.86	0.90
60	20	0.01	0.00	0.07	0.03	0.03	0.00	0.50	0.45	0.50	0.50	0.51	0.85	0.77	0.85	0.85	0.90
60	20	0.1	0.00	0.14	0.09	0.10	0.01	0.47	0.38	0.45	0.44	0.51	0.76	0.63	0.72	0.72	0.87
60	20	0.3	-0.02	0.22	0.23	0.22	0.04	0.40	0.26	0.29	0.29	0.47	0.66	0.44	0.46	0.48	0.80
60	50	0.001	0.00	0.04	0.01	0.01	0.00	0.50	0.46	0.50	0.49	0.50	0.89	0.82	0.89	0.88	0.90
60	50	0.01	0.00	0.06	0.02	0.02	0.00	0.50	0.45	0.49	0.49	0.50	0.88	0.80	0.87	0.87	0.90
60	50	0.1	0.00	0.13	0.11	0.11	0.01	0.48	0.40	0.43	0.43	0.52	0.78	0.65	0.69	0.69	0.88
60	50	0.3	-0.02	0.21	0.25	0.25	0.04	0.41	0.27	0.27	0.27	0.49	0.67	0.45	0.43	0.44	0.81
50% missing																	
24	20	0.001	0.00	0.38	0.25	0.25	0.15	0.37	0.02	0.34	0.36	0.46	0.86	0.47	0.77	0.76	0.70
24	20	0.01	-0.01	0.39	0.26	0.25	0.16	0.36	0.01	0.33	0.34	0.45	0.83	0.43	0.72	0.72	0.68
24	20	0.1	-0.06	0.36	0.27	0.27	0.19	0.32	-0.02	0.26	0.26	0.36	0.68	0.26	0.53	0.57	0.57
24	20	0.3	-0.19	0.31	0.27	0.27	0.18	0.32	-0.05	0.20	0.20	0.32	0.61	0.13	0.36	0.41	0.50
24	50	0.001	0.00	0.39	0.25	0.24	0.14	0.41	0.04	0.37	0.36	0.50	0.91	0.51	0.81	0.83	0.72
24	50	0.01	0.00	0.39	0.24	0.24	0.13	0.39	0.02	0.35	0.34	0.49	0.89	0.49	0.79	0.8	0.73
24	50	0.1	-0.05	0.36	0.27	0.27	0.18	0.31	-0.02	0.24	0.25	0.37	0.69	0.28	0.55	0.58	0.59
24	50	0.3	-0.17	0.32	0.29	0.28	0.18	0.31	-0.08	0.18	0.20	0.32	0.60	0.13	0.35	0.41	0.51
60	20	0.001	0.00	0.44	0.27	0.27	0.04	0.37	-0.07	0.35	0.36	0.57	0.86	0.41	0.8	0.77	0.84
60	20	0.01	-0.01	0.45	0.27	0.27	0.06	0.36	-0.10	0.34	0.34	0.55	0.82	0.36	0.76	0.74	0.81
60	20	0.1	-0.06	0.44	0.27	0.27	0.10	0.32	-0.15	0.28	0.29	0.47	0.67	0.16	0.58	0.58	0.71
60	20	0.3	-0.18	0.41	0.28	0.29	0.11	0.30	-0.20	0.21	0.21	0.40	0.59	0.01	0.42	0.42	0.64
60	50	0.001	0.00	0.45	0.27	0.27	0.01	0.41	-0.04	0.37	0.38	0.62	0.9	0.45	0.81	0.80	0.88
60	50	0.01	0.00	0.45	0.27	0.27	0.02	0.38	-0.06	0.35	0.35	0.61	0.89	0.44	0.81	0.79	0.87
60	50	0.1	-0.04	0.45	0.29	0.29	0.11	0.31	-0.15	0.26	0.26	0.45	0.67	0.18	0.57	0.56	0.70
60	50	0.3	-0.17	0.43	0.30	0.30	0.12	0.30	-0.22	0.19	0.19	0.40	0.58	-0.02	0.38	0.39	0.63

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 5: Estimates of the intracluster correlation coefficient and residual variance based on methods to handle 25% and 50% missing data in categorical cluster level covariate

No. clusters	Cluster size	ICC	MM	MD	SL	FE	AG	σ^2	MM	MD	SL	FE	AG
25% missing													
24	20	0.001	0.00	0.05	0.02	0.03	0.01	0.749	0.74	0.75	0.73	0.73	0.74
24	20	0.01	0.01	0.06	0.04	0.04	0.02	0.74	0.74	0.74	0.73	0.73	0.74
24	20	0.1	0.08	0.18	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
24	20	0.3	0.22	0.40	0.35	0.35	0.34	0.45	0.45	0.45	0.44	0.44	0.45
24	50	0.001	0.00	0.03	0.02	0.02	0.00	0.749	0.75	0.75	0.74	0.74	0.75
24	50	0.01	0.01	0.04	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
24	50	0.1	0.08	0.17	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
24	50	0.3	0.23	0.39	0.35	0.35	0.33	0.45	0.45	0.45	0.44	0.44	0.45
60	20	0.001	0.00	0.04	0.02	0.02	0.00	0.749	0.75	0.75	0.74	0.74	0.75
60	20	0.01	0.01	0.05	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
60	20	0.1	0.08	0.17	0.14	0.14	0.11	0.65	0.65	0.65	0.64	0.64	0.65
60	20	0.3	0.22	0.39	0.35	0.35	0.32	0.45	0.45	0.45	0.44	0.44	0.45
60	50	0.001	0.00	0.03	0.01	0.01	0.00	0.749	0.75	0.75	0.74	0.74	0.75
60	50	0.01	0.01	0.05	0.03	0.03	0.01	0.74	0.74	0.74	0.73	0.73	0.74
60	50	0.1	0.08	0.17	0.14	0.14	0.10	0.65	0.65	0.65	0.64	0.64	0.65
60	50	0.3	0.22	0.39	0.35	0.35	0.31	0.45	0.45	0.45	0.44	0.44	0.45
50% missing													
24	20	0.001	0.01	0.13	0.08	0.09	0.07	0.749	0.75	0.75	0.71	0.71	0.75
24	20	0.01	0.01	0.14	0.09	0.10	0.08	0.74	0.74	0.74	0.70	0.70	0.74
24	20	0.1	0.07	0.22	0.17	0.19	0.18	0.65	0.65	0.65	0.62	0.62	0.65
24	20	0.3	0.19	0.41	0.35	0.37	0.38	0.45	0.45	0.45	0.43	0.43	0.45
24	50	0.001	0.00	0.13	0.08	0.09	0.06	0.749	0.75	0.75	0.71	0.70	0.75
24	50	0.01	0.01	0.14	0.08	0.10	0.07	0.74	0.74	0.74	0.70	0.70	0.74
24	50	0.1	0.07	0.22	0.17	0.19	0.18	0.65	0.65	0.65	0.62	0.62	0.65
24	50	0.3	0.20	0.41	0.35	0.37	0.38	0.45	0.45	0.45	0.43	0.43	0.45
60	20	0.001	0.00	0.13	0.09	0.09	0.03	0.749	0.75	0.75	0.71	0.72	0.75
60	20	0.01	0.01	0.14	0.10	0.10	0.04	0.74	0.74	0.74	0.71	0.71	0.74
60	20	0.1	0.07	0.23	0.18	0.19	0.14	0.65	0.65	0.65	0.62	0.62	0.65
60	20	0.3	0.20	0.40	0.36	0.37	0.34	0.45	0.45	0.45	0.43	0.43	0.45
60	50	0.001	0.00	0.13	0.08	0.09	0.02	0.749	0.75	0.75	0.71	0.71	0.75
60	50	0.01	0.01	0.14	0.09	0.10	0.03	0.74	0.74	0.74	0.70	0.70	0.74
60	50	0.1	0.07	0.22	0.18	0.18	0.14	0.65	0.65	0.65	0.63	0.63	0.65
60	50	0.3	0.20	0.41	0.37	0.37	0.35	0.45	0.45	0.45	0.43	0.43	0.45

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

Table 6: Coverage of true values by the 95% confidence interval of the regression coefficients based on methods to handle 25% and 50% missing data in continuous cluster level covariate

No. clusters	Cluster size	ICC	$\beta_0 = 0$					$\beta_1 = 0.5$					$\beta_2 = 0.9$				
			MM	MD	SL	FE	AG	MM	MD	SL	FE	AG	MM	MD	SL	FE	AG
25% missing																	
24	20	0.001	95	82	96	96	96	95	93	97	97	96	96	93	98	97	96
24	20	0.01	94	82	95	95	96	94	94	97	97	96	95	92	98	97	96
24	20	0.1	92	75	88	89	94	90	88	94	93	94	89	84	90	90	96
24	20	0.3	89	72	74	75	92	91	83	84	85	93	84	77	59	65	95
24	50	0.001	94	85	97	96	95	95	96	98	97	96	95	95	97	97	94
24	50	0.01	92	84	96	95	94	93	94	97	97	95	92	93	96	96	94
24	50	0.1	89	75	83	82	94	92	89	90	87	94	88	85	74	77	95
24	50	0.3	88	69	61	64	91	89	83	67	66	94	85	77	32	38	94
60	20	0.001	96	68	96	96	96	96	81	98	98	96	92	72	97	97	95
60	20	0.01	94	66	94	94	95	94	83	98	98	95	91	73	97	97	94
60	20	0.1	93	56	81	80	95	93	67	94	93	94	79	56	72	73	95
60	20	0.3	92	49	48	50	92	88	57	64	65	92	71	47	23	22	93
60	50	0.001	95	72	97	96	95	95	87	98	98	95	95	81	98	98	95
60	50	0.01	93	68	95	94	94	93	85	97	97	94	93	78	96	96	94
60	50	0.1	93	57	68	68	95	92	70	88	86	93	81	60	47	48	95
60	50	0.3	92	49	31	32	92	88	56	41	40	91	74	47	3	4	93
50% missing																	
24	20	0.001	95	16	34	36	76	86	44	90	82	96	91	81	93	90	93
24	20	0.01	94	16	34	38	74	85	42	89	82	97	84	74	91	85	94
24	20	0.1	88	25	48	49	74	78	47	80	73	95	68	59	72	70	91
24	20	0.3	78	34	56	61	83	82	53	66	68	96	74	58	51	57	92
24	50	0.001	95	12	10	17	74	85	42	84	66	96	96	92	92	83	93
24	50	0.01	92	12	15	20	77	80	43	79	63	97	91	89	89	81	93
24	50	0.1	88	23	33	37	76	77	44	53	53	95	71	62	54	55	91
24	50	0.3	82	34	47	51	81	83	52	49	51	96	73	57	37	41	92
60	20	0.001	95	4	3	2	86	66	4	73	73	74	89	60	91	85	94
60	20	0.01	94	4	3	4	81	62	4	68	67	78	83	49	85	80	91
60	20	0.1	86	8	14	14	79	71	7	58	56	87	66	25	59	53	86
60	20	0.3	69	14	27	27	77	80	12	46	43	87	68	23	37	32	86
60	50	0.001	96	2	0	0	97	58	2	61	59	46	94	66	87	79	98
60	50	0.01	92	2	0	0	94	51	3	49	51	59	91	67	84	77	97
60	50	0.1	90	8	2	3	74	63	8	29	32	86	65	27	37	37	84
60	50	0.3	69	13	14	16	78	81	12	32	27	87	67	20	21	16	85

Abbreviations: ICC = intracluster correlation coefficient, MM = mixed model, MD = mode substitution, SL = single level multiple imputation (MI), FE = fixed effects MI, AG = MI aggregate imputation.

5 Discussion

We performed a simulation study to evaluate strategies to handle missing cluster level covariates under the MAR assumption. We varied the total number of clusters, cluster size, ICC, percentage of missingness, and examined the methods when the cluster level covariate was continuous and categorical. We evaluated bias among the fixed and random parameter estimates, as well as coverage of the true values of the fixed parameters.

For both continuous and categorical cluster level covariates, the mixed model produced better estimates of the regression coefficients and ICC when the ICC was low (ICC $\leq$ 0.1) and with 25% missing cluster level data. However, when the ICC was higher (ICC $>$ 0.1), the mixed model generated severely biased estimates of the regression coefficients, which became worse when the percentage of missing data increased to 50%. MI aggregate imputation performed best when the ICC was higher (ICC $>$ 0.1), though it tended to overestimate the ICC similar to the other imputation approaches. When the cluster level covariate was categorical, none of the strategies performed well when 50% of the data were missing. Overall, the worst methods for the missing continuous and categorical cluster level covariate were mean substitution and mode substitution, respectively. Although these single imputation methods are simple, they should not be used, because they produced biased fixed and random effect estimates, especially when the cluster level covariate was a categorical variable.

In the context of longitudinal data, Cheung [13] compared methods to handle time-invariant data under MCAR, and found the complete case analysis and mixed model to perform best. He also concluded mean substitution to be acceptable when the amount of missing data was low (10%). Among educational data, Gibson et al. [12] investigated missing continuous cluster level data under MCAR, and found that complete case analysis and mean substitution performed similarly, but that complete case analysis was the only method to perform well when estimating random effects with a higher percentage of missing data (40%) and smaller total clusters ($N = 30$). However, MCAR may be too strong of an assumption in scenarios such as cluster randomized trials, since the reason for missingness may depend on the treatment or other observed data. For this reason, appropriate methods to handle missing multilevel data should be chosen based on reasonable assumptions about the missing data.

Van Buuren studied missing data strategies to handle missing outcomes and covariates at the individual level under MAR. He compared the mixed model, single level MI ignoring clustering, fixed effects MI, and multilevel MI, which incorporates clustering into the imputation process via the Gibbs sampler. He found complete case analysis to be a poor strategy when missingness occurred among individual level covariates, and concluded multilevel MI to be the best strategy overall for missing individual level data [9]. However, multilevel MI as implemented in the `mice` package in R is unable to impute missing cluster

level variables.

In general, we found that neither single level MI nor fixed effects MI performed well when imputing cluster level variables, particularly when the percentage of missing data was higher. Both methods produced biased fixed and random parameter estimates, and were undercoveraged for the fixed parameters. Van Buuren found fixed effects MI to perform reasonably well when imputing individual level covariates, though this method had computation problems when the cluster size was small ($n_i = 20$) due to the low number of observations per cluster (≤ 3) after generating missing data. He found single level MI ignoring clustering to be less successful in generating appropriate fixed and random parameter estimates [9]. Single level MI is still generally used in practice [5], even though it has been shown to perform well only under the restrictive assumption that the continuous individual level outcome is MCAR and the ICC is very low (ICC < 0.005) [7]. Single level MI should not be used to impute cluster level covariates because it has been shown to perform poorly, regardless of whether the data are MCAR [12, 13] or MAR.

A strength of our study is that we assessed missing data strategies under the MAR assumption, which may be a more likely scenario in practice. Along with the continuous cluster level covariate, we also investigated the behavior of methods when the cluster level covariate was categorical, which may be more widely found among cluster level data. We evaluated scenarios with small and large total number of clusters, cluster sizes, and ICC. We did not examine the scenario in which missing data occur in multiple levels within the hierarchical structure. For example, along with missing cluster level covariates, missing data can occur among individual level outcomes and covariates simultaneously. This scenario is perhaps most commonly seen in practice. Van Buuren studied the scenario in which missing data occur among individual level outcomes and covariates simultaneously, and found multilevel MI to perform best, though still not ideal [9]. Another, more complex setting to examine is a three-level model, such as a longitudinal cluster randomized trial, which includes clusters, individuals per cluster, and measurements per individual. Further investigation of appropriate approaches when missing data occur in different places among clustered data is needed. Although it is possible to test between MCAR and MAR, it is not possible to test between MAR and MNAR since the data are missing. For this reason, it would be beneficial to examine the performance of approaches under departures from the MAR assumption in future.

Based on our simulations, we recommend using the mixed model for missing cluster level covariates when the ICC is small ($\rho \leq 0.1$) and the percentage of missing data is low ($\leq 25\%$), as long as there are a large number of clusters. Otherwise, MI aggregate imputation should be used to impute missing cluster level covariates, though caution should be taken if the percentage of missing cluster level covariates is high. Mean and mode substitution are not recommended as effective strategies to imputed missing cluster level covariates.

References

[1] Jerome Cornfield. Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108(2):100–102, 1978.

[2] Marion K Campbell, Jill Mollison, Nick Steen, Jeremy M Grimshaw, and Martin Eccles. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice*, 17(2):192–196, 2000.

[3] Melanie L Bell and Joanne E McKenzie. Designing psycho-oncology randomised trials and cluster randomised trials: variance components and intra-cluster correlation of commonly used psychosocial measures. *Psycho-Oncology*, 22(8):1738–1747, 2013.

[4] Geoffrey Adams, Martin C Gulliford, Obioha C Ukoumunne, Sandra Eldridge, Susan Chinn, and Michael J Campbell. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57(8):785–794, 2004.

[5] Mallorie H. Fiero, Shuang Huang, Eyal Oren, and Melanie L. Bell. Statistical analysis and handling of missing data in cluster randomised trials: a systematic review. *Trials*, 17:72, 2015.

[6] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[7] Monica Taljaard, Allan Donner, and Neil Klar. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, 50(3):329–345, 2008.

[8] Rebecca R Andridge. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal*, 53(1):57–74, 2011.

[9] Stef Van Buuren et al. *Multiple imputation of multilevel data*. Routledge New York, NY, 2011.

[10] Jinhui Ma, Noori Akhtar-Danesh, Lisa Dolovich, and Lehana Thabane. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology*, 11(1):1, 2011.

[11] Jinhui Ma, P Raina, J Beyene, and L Thabane. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat*, 2:93–103, 2012.

[12] Nicole Morgan Gibson and Stephen Olejnik. Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, 63(2):204–238, 2003.

[13] Mike W-L Cheung. Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random. *Organizational Research Methods*, 2007.

[14] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[15] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[16] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011.

[17] Jaap Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. PhD thesis, Erasmus University Rotterdam, 1999.

[18] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 1987.

[19] S Van Buuren and CGM Oudshoorn. mice: Multivariate imputation by chained equations. r package version 1.16, 2007.

APPENDIX E: SUPPLEMENTARY FILE 1 - SEARCH STRATEGY

Supplementary file 1

Search terms and strategy used in PubMed. The same search was also performed in Web of Science (all databases) and PsycINFO.

Cluster randomized OR cluster randomised OR community trial OR community randomized OR community randomised OR group randomized OR group randomised OR (cluster AND trial) Limiters: all in title or abstract, August 1, 2013 – July 31, 2014 1285 articles found

APPENDIX F: SUPPLEMENTARY FILE 2 - DATA EXTRACTION

Supplementary file 2

Specific details on data items, including relevant coding used during the data extraction process.

Data items*

1. Year
2. Month
3. Journal
4. Author
    a. Last name of first author
5. Stepped wedge
    a. Yes, No
6. Pilot/feasibility
    a. Yes, No
7. If pilot/feasibility, were hypothesis tests performed?
    a. Yes, No, NA
8. If pilot/feasibility, were feasibility outcomes stated?
    a. Yes, No, NA
9. Outcome
10. Type of outcome
    a. Binary, Continuous, Count
11. How often outcome was collected at individual level
    a. Single, Repeated
12. How outcome was treated in the primary analysis
    a. Single, Repeated
13. Unit of randomization
    a. E.g. clinic, practitioner
14. Stratification/Matching/Minimization in randomization
    a. Stratification, Matching, Minimization, No
15. No. clusters randomized
16. No. clusters missing outcome
17. % missing - cluster level
18. Total no. participants randomized
19. No. participants missing outcome
20. % missing - individual level
21. If survey data, response rate at time period of primary analysis
22. Average no. participants per cluster
23. Min no. participants in cluster
24. Max no. participants in cluster
25. Presented sample size calculation
    a. Yes, No
26. Accounted for clustering in sample size
    a. Yes, No
27. Reported ICC or CV in sample size

28. Accounted for missing outcome data in calculation
    a. Yes, No
29. If yes, accounted missingness clusters and/or individuals
    a. Clusters, Individuals, Both, Unclear
30. Reported attrition rate in sample size
31. Primary analysis
32. Clustering accounted for in analysis
    a. Yes, No
33. Observed ICC or CV reported (primary outcome)
34. If so, how does it compare to ICC or CV used in sample size calculation?
    a. 100 * (Observed ICC – Sample size ICC) / Sample size ICC
35. GEE correction
    a. Yes, No, NA
36. If yes, what type?
    a. Bias correction, DF adjustment, Bootstrap
37. Method missing data in primary analysis
    a. Complete case, single imputation (LOCF, worst case, etc.), multiple imputation, mixed model, GEE, GEE IPW, Bayesian, Unclear
38. If imputation, was it multilevel?
    a. Yes, No, NA, Unclear
39. Sensitivity analysis
    a. Complete case, single imputation (LOCF, worst case, etc.), multiple imputation, mixed model, GEE, GEE IPW, Bayesian, No, Unclear
40. Level of reporting sensitivity analysis
    a. Sentence, Paragraph, Tabulation, NA
41. Notes

* If any item is not applicable, not reported or unclear, indicate “NA”, “NR” or “Unclear”, respectively, in appropriate field.

APPENDIX G: ADDITIONAL FILE 1 - PRISMA CHECKLIST

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	6
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6-7
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6 (Protocol)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	8-10
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	8-9
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	10, Figure 1
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	10-11
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	NA
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	13-14
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14, 17-18
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16-17
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	14-16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

APPENDIX H: ADDITIONAL FILE 2 - REFERENCES OF 86 INCLUDED TRIALS IN REVIEW

Supplemental File

Below are the references of the 86 trials included in the review.

1. Arvidsson H, Olin E, Strand J, et al. Effects of the two-way communication checklist (2-COM): a one-year cluster randomized study in a group of severely mentally ill persons. Int J Soc Psychiatry 2014;**60**(1):95-102 doi: 10.1177/0020764012467145.
2. Baatjies R, Meijster T, Heederik D, et al. Effectiveness of interventions to reduce flour dust exposures in supermarket bakeries in South Africa. Occup Environ Med 2014;**71**(12):811-8 doi: 10.1136/oemed-2013-101971.
3. Bavarian N, Lewis KM, Dubois DL, et al. Using social-emotional and character development to improve academic outcomes: a matched-pair, cluster-randomized controlled trial in low-income, urban schools. J Sch Health 2013;**83**(11):771-9 doi: 10.1111/josh.12093.
4. Bird C, Ame S, Albonico M, et al. Do shoes reduce hookworm infection in school-aged children on Pemba Island, Zanzibar? A pragmatic trial. Trans R Soc Trop Med Hyg 2014;**108**(5):297-304 doi: 10.1093/trstmh/tru037.
5. Campbell L, Novak I, McIntyre S, et al. A KT intervention including the evidence alert system to improve clinician's evidence-based practice behavior--a cluster randomized controlled trial. Implement Sci 2013;**8**:132 doi: 10.1186/1748-5908-8-132.
6. Carroll AE, Bauer NS, Dugan TM, et al. Use of a computerized decision aid for developmental surveillance and screening: a randomized clinical trial. JAMA Pediatr 2014;**168**(9):815-21 doi: 10.1001/jamapediatrics.2014.464.
7. Chan SS, Leung DY, Leung AY, et al. A nurse-delivered brief health education intervention to improve pneumococcal vaccination rate among older patients with chronic diseases: A cluster randomized controlled trial. Int J Nurs Stud 2015;**52**(1):317-24 doi: 10.1016/j.ijnurstu.2014.06.008.
8. Cohen KE, Morgan PJ, Plotnikoff RC, et al. Physical Activity and Skills Intervention: SCORES Cluster Randomized Controlled Trial. Med Sci Sports Exerc 2014 doi: 10.1249/MSS.0000000000000452.
9. Colon-Emeric CS, McConnell E, Pinheiro SO, et al. CONNECT for better fall prevention in nursing homes: results from a pilot intervention study. J Am Geriatr Soc 2013;**61**(12):2150-9 doi: 10.1111/jgs.12550.
10. Connor CM, Morrison FJ, Fishman B, et al. A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. Psychol Sci 2013;**24**(8):1408-19 doi: 10.1177/0956797612472204.
11. Craig T, Shepherd G, Rinaldi M, et al. Vocational rehabilitation in early psychosis: cluster randomised trial. Br J Psychiatry 2014;**205**(2):145-50 doi: 10.1192/bjp.bp.113.136283.
12. de Graaff JC, Cuper NJ, Mungra RA, et al. Near-infrared light to aid peripheral intravenous cannulation in children: a cluster randomised clinical trial of three devices. Anaesthesia 2013;**68**(8):835-45 doi: 10.1111/anae.12294.
13. Deales A, Fratini M, Romano S, et al. Care manager to control cardiovascular risk factors in primary care: the Raffaello cluster randomized trial. Nutr Metab Cardiovasc Dis 2014;**24**(5):563-71 doi: 10.1016/j.numecd.2013.11.008.
14. Deressa W, Yihdego YY, Kebede Z, et al. Effect of combining mosquito repellent and insecticide treated net on malaria prevalence in Southern Ethiopia: a cluster-randomised trial. Parasit Vectors 2014;**7**:132 doi: 10.1186/1756-3305-7-132.

15. Dixon A, Clarkin C, Barrowman N, et al. Reduction of radial-head subluxation in children by triage nurses in the emergency department: a cluster-randomized controlled trial. CMAJ 2014;**186**(9):E317-23 doi: 10.1503/cmaj.131101.
16. Duda JL, Williams GC, Ntoumanis N, et al. Effects of a standard provision versus an autonomy supportive exercise referral programme on physical activity, quality of life and well-being indicators: a cluster randomised controlled trial. Int J Behav Nutr Phys Act 2014;**11**:10 doi: 10.1186/1479-5868-11-10.
17. Ebenezer R, Gunawardena K, Kumarendran B, et al. Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of schoolchildren in Sri Lanka's plantation sector. Trop Med Int Health 2013;**18**(8):942-51 doi: 10.1111/tmi.12128.
18. Fink G, Robyn PJ, Sié A, et al. Does health insurance improve health?: Evidence from a randomized community-based insurance rollout in rural Burkina Faso. J Health Econ 2013;**32**(6):1043-56 doi: 10.1016/j.jhealeco.2013.08.003.
19. Flax VL, Negerie M, Ibrahim AU, et al. Integrating group counseling, cell phone messaging, and participant-generated songs and dramas into a microcredit program increases Nigerian women's adherence to international breastfeeding recommendations. J Nutr 2014;**144**(7):1120-4 doi: 10.3945/jn.113.190124.
20. Freiberger E, Blank WA, Salb J, et al. Effects of a complex intervention on fall risk in the general practitioner setting: a cluster randomized controlled trial. Clin Interv Aging 2013;**8**:1079-88 doi: 10.2147/CIA.S46218.
21. Fuller JM, Wong KK, Grunstein R, et al. A comparison of screening methods for sleep disorders in Australian community pharmacies: a randomized controlled trial. PLoS One 2014;**9**(6):e101003 doi: 10.1371/journal.pone.0101003.
22. Galik E, Resnick B, Hammersla M, et al. Optimizing function and physical activity among nursing home residents with dementia: testing the impact of function-focused care. Gerontologist 2014;**54**(6):930-43 doi: 10.1093/geront/gnt108.
23. Gärtner FR, Nieuwenhuijsen K, Ketelaar SM, et al. The mental vitality @ work study: effectiveness of a mental module for workers' health surveillance for nurses and allied health care professionals on their help-seeking behavior. J Occup Environ Med 2013;**55**(10):1219-29 doi: 10.1097/JOM.0b013e31829f310a.
24. Haller DM, Meynard A, Lefebvre D, et al. Effectiveness of training family physicians to deliver a brief intervention to address excessive substance use among young patients: a cluster randomized controlled trial. CMAJ 2014;**186**(8):E263-72 doi: 10.1503/cmaj.131301.
25. Hamid S, Dunsiger S, Seiden A, et al. Impact of a diabetes control and management intervention on health care utilization in American Samoa. Chronic Illn 2014;**10**(2):122-34 doi: 10.1177/1742395313502367.
26. Haugen AS, Søfteland E, Almeland SK, et al. Effect of the World Health Organization Checklist on Patient Outcomes: A Stepped Wedge Cluster Randomized Controlled Trial. Ann Surg 2014 doi: 10.1097/SLA.0000000000000716.
27. Heyland DK, Murch L, Cahill N, et al. Enhanced protein-energy provision via the enteral route feeding protocol in critically ill patients: results of a cluster randomized trial. Crit Care Med 2013;**41**(12):2743-53 doi: 10.1097/CCM.0b013e31829efef5.
28. Hiemstra M, Ringlever L, Otten R, et al. Long-term effects of a home-based smoking prevention program on smoking initiation: a cluster randomized controlled trial. Prev Med 2014;**60**:65-70 doi: 10.1016/j.ypmed.2013.12.012.
29. Hirani SP, Beynon M, Cartwright M, et al. The effect of telecare on the quality of life and psychological well-

being of elderly recipients of social care over a 12-month period: the Whole Systems Demonstrator cluster randomised trial. Age Ageing 2014;**43**(3):334-41 doi: 10.1093/ageing/aft185.

30. Inauen J, Tobias R, Mosler HJ. The role of commitment strength in enhancing safe water consumption: mediation analysis of a cluster-randomized trial. Br J Health Psychol 2014;**19**(4):701-19 doi: 10.1111/bjhp.12068.
31. Isensee B, Hansen J, Maruska K, et al. Effects of a school-based prevention programme on smoking in early adolescence: a 6-month follow-up of the 'Eigenstandig werden' cluster randomised trial. BMJ Open 2014;**4**(1):e004422 doi: 10.1136/bmjopen-2013-004422.
32. Ismail KM, Kettle C, Macdonald SE, et al. Perineal Assessment and Repair Longitudinal Study (PEARLS): a matched-pair cluster randomized trial. BMC Med 2013;**11**:209 doi: 10.1186/1741-7015-11-209.
33. Kauye F, Jenkins R, Rahman A. Training primary health care workers in mental health and its impact on diagnoses of common mental disorders in primary care of a developing country, Malawi: a cluster-randomized controlled trial. Psychol Med 2014;**44**(3):657-66 doi: 10.1017/S0033291713001141.
34. Ketelaar SM, Nieuwenhuijsen K, Gärtner FR, et al. Effect of an E-mental health approach to workers' health surveillance versus control group on work functioning of hospital employees: a cluster-RCT. PLoS One 2013;**8**(9):e72546 doi: 10.1371/journal.pone.0072546.
35. Lerner-Geva L, Bar-Zvi E, Levitan G, et al. An intervention for improving the lifestyle habits of kindergarten children in Israel: a cluster-randomised controlled trial investigation. Public Health Nutr 2014:1-8 doi: 10.1017/S136898001400024X.
36. Little P, Stuart B, Francis N, et al. Effects of internet-based training on antibiotic prescribing rates for acute respiratory-tract infections: a multinational, cluster, randomised, factorial, controlled trial. Lancet 2013;**382**(9899):1175-82 doi: 10.1016/S0140-6736(13)60994-0.
37. Madigan SM, Fleming P, Wright ME, et al. A cluster randomised controlled trial of a nutrition education intervention in the community. J Hum Nutr Diet 2014;**27 Suppl 2**:12-20 doi: 10.1111/jhn.12079.
38. McCrow J, Sullivan KA, Beattie ER. Delirium knowledge and recognition: a randomized controlled trial of a web-based educational intervention for acute care nurses. Nurse Educ Today 2014;**34**(6):912-7 doi: 10.1016/j.nedt.2013.12.006.
39. Meeks S, Van Haitsma K, Schoenbachler B, et al. BE-ACTIV for Depression in Nursing Homes: Primary Outcomes of a Randomized Clinical Trial. J Gerontol B Psychol Sci Soc Sci 2015;**70**(1):13-23 doi: 10.1093/geronb/gbu026.
40. Menchetti M, Sighinolfi C, Di Michele V, et al. Effectiveness of collaborative care for depression in Italy. A randomized controlled trial. Gen Hosp Psychiatry 2013;**35**(6):579-86 doi: 10.1016/j.genhosppsych.2013.07.009.
41. Mengistie B, Berhane Y, Worku A. Household water chlorination reduces incidence of diarrhea among under-five children in rural Ethiopia: a cluster randomized controlled trial. PLoS One 2013;**8**(10):e77887 doi: 10.1371/journal.pone.0077887.
42. Meyer U, Schindler C, Zahner L, et al. Long-term effect of a school-based physical activity program (KISS) on fitness and adiposity in children: a cluster-randomized controlled trial. PLoS One 2014;**9**(2):e87929 doi: 10.1371/journal.pone.0087929.
43. Muhumuza S, Olsen A, Katahoire A, et al. Effectiveness of a pre-treatment snack on the uptake of mass treatment for schistosomiasis in Uganda: a cluster randomized trial. PLoS Med 2014;**11**(5):e1001640 doi: 10.1371/journal.pmed.1001640.

44. Na JU, Lee TR, Kang MJ, et al. Basic life support skill improvement with newly designed renewal programme: cluster randomised study of small-group-discussion method versus practice-while-watching method. Emerg Med J 2014;**31**(12):964-9 doi: 10.1136/emermed-2013-202379.
45. Nauta J, Knol DL, Adriaensens L, et al. Prevention of fall-related injuries in 7-year-old to 12-year-old children: a cluster randomised controlled trial. Br J Sports Med 2013;**47**(14):909-13 doi: 10.1136/bjsports-2012-091439.
46. Ochola SA, Labadarios D, Nduati RW. Impact of counselling on exclusive breast-feeding practices in a poor urban setting in Kenya: a randomized controlled trial. Public Health Nutr 2013;**16**(10):1732-40 doi: 10.1017/S1368980012004405.
47. Palmu AA, Jokinen J, Nieminen H, et al. Effect of pneumococcal Haemophilus influenzae protein D conjugate vaccine (PHiD-CV10) on outpatient antimicrobial purchases: a double-blind, cluster randomised phase 3-4 trial. Lancet Infect Dis 2014;**14**(3):205-12 doi: 10.1016/S1473-3099(13)70338-4.
48. Papish A, Kassam A, Modgill G, et al. Reducing the stigma of mental illness in undergraduate medical education: a randomized controlled trial. BMC Med Educ 2013;**13**:141 doi: 10.1186/1472-6920-13-141.
49. Pasha O, McClure EM, Wright LL, et al. A combined community- and facility-based approach to improve pregnancy outcomes in low-resource settings: a Global Network cluster randomized trial. BMC Med 2013;**11**:215 doi: 10.1186/1741-7015-11-215.
50. Penfold S, Manzi F, Mkumbo E, et al. Effect of home-based counselling on newborn care practices in southern Tanzania one year after implementation: a cluster-randomised controlled trial. BMC Pediatr 2014;**14**:187 doi: 10.1186/1471-2431-14-187.
51. Power M, Tyrrell PJ, Rudd AG, et al. Did a quality improvement collaborative make stroke care better? A cluster randomized trial. Implement Sci 2014;**9**(1):40 doi: 10.1186/1748-5908-9-40.
52. Primack BA, Douglas EL, Land SR, et al. Comparison of media literacy and usual education to prevent tobacco use: a cluster-randomized trial. J Sch Health 2014;**84**(2):106-15 doi: 10.1111/josh.12130.
53. Rat C, Quereux G, Riviere C, et al. Targeted melanoma prevention intervention: a cluster randomized controlled trial. Ann Fam Med 2014;**12**(1):21-8 doi: 10.1370/afm.1600.
54. Reynolds GS, Bennett JB. A cluster randomized trial of alcohol prevention in small businesses: a cascade model of help seeking and risk reduction. Am J Health Promot 2015;**29**(3):182-91 doi: 10.4278/ajhp.121212-QUAN-600.
55. Richards DA, Hill JJ, Gask L, et al. Clinical effectiveness of collaborative care for depression in UK primary care (CADET): cluster randomised controlled trial. BMJ 2013;**347**:f4913
56. Richter L, Rotheram-Borus MJ, Van Heerden A, et al. Pregnant women living with HIV (WLH) supported at clinics by peer WLH: a cluster randomized controlled trial. AIDS Behav 2014;**18**(4):706-15 doi: 10.1007/s10461-014-0694-2.
57. Saboori S, Greene LE, Moe CL, et al. Impact of regular soap provision to primary schools on hand washing and E. coli hand contamination among pupils in Nyanza Province, Kenya: a cluster-randomized trial. Am J Trop Med Hyg 2013;**89**(4):698-708 doi: 10.4269/ajtmh.12-0387.
58. Santos RG, Durksen A, Rabbanni R, et al. Effectiveness of peer-based healthy living lesson plans on anthropometric measures and physical activity in elementary school students: a cluster randomized trial. JAMA Pediatr 2014;**168**(4):330-7 doi: 10.1001/jamapediatrics.2013.3688.
59. Shakeshaft A, Doran C, Petrie D, et al. The effectiveness of community action in reducing risky alcohol consumption and harm: a cluster randomised controlled trial. PLoS Med 2014;**11**(3):e1001617 doi:

10.1371/journal.pmed.1001617.

60. Smidth M, Olesen F, Fenger-Grøn M, et al. Patient-experienced effect of an active implementation of a disease management programme for COPD - a randomised trial. BMC Fam Pract 2013;**14**:147 doi: 10.1186/1471-2296-14-147.
61. Snow PC, Eadie PA, Connell J, et al. Oral language supports early literacy: a pilot cluster randomized trial in disadvantaged schools. Int J Speech Lang Pathol 2014;**16**(5):495-506 doi: 10.3109/17549507.2013.845691.
62. Sorensen G, Pednekar MS, Sinha DN, et al. Effects of a tobacco control intervention for teachers in India: results of the Bihar school teachers study. Am J Public Health 2013;**103**(11):2035-40 doi: 10.2105/AJPH.2013.301303.
63. Stallard P, Phillips R, Montgomery AA, et al. A cluster randomised controlled trial to determine the clinical effectiveness and cost-effectiveness of classroom-based cognitive-behavioural therapy (CBT) in reducing symptoms of depression in high-risk adolescents. Health Technol Assess 2013;**17**(47):vii-xvii, 1-109 doi: 10.3310/hta17470.
64. Stanton CK, Newton S, Mullany LC, et al. Effect on postpartum hemorrhage of prophylactic oxytocin (10 IU) by injection by community health officers in Ghana: a community-based, cluster-randomized trial. PLoS Med 2013;**10**(10):e1001524 doi: 10.1371/journal.pmed.1001524.
65. Svarstad BL, Kotchen JM, Shireman TI, et al. Improving refill adherence and hypertension control in black patients: Wisconsin TEAM trial. J Am Pharm Assoc (2003) 2013;**53**(5):520-9 doi: 10.1331/JAPhA.2013.12246.
66. Taddio A, Smart S, Sheedy M, et al. Impact of prenatal education on maternal utilization of analgesic interventions at future infant vaccinations: a cluster randomized trial. Pain 2014;**155**(7):1288-92 doi: 10.1016/j.pain.2014.03.024.
67. Tannenbaum C, Agnew R, Benedetti A, et al. Effectiveness of continence promotion for older women via community organisations: a cluster randomised trial. BMJ Open 2013;**3**(12):e004135 doi: 10.1136/bmjopen-2013-004135.
68. Tannenbaum C, Martin P, Tamblyn R, et al. Reduction of inappropriate benzodiazepine prescriptions among older adults through direct patient education: the EMPOWER cluster randomized trial. JAMA Intern Med 2014;**174**(6):890-8 doi: 10.1001/jamainternmed.2014.949.
69. Tine RC, Ndour CT, Faye B, et al. Feasibility, safety and effectiveness of combining home based malaria management and seasonal malaria chemoprevention in children less than 10 years in Senegal: a cluster-randomised trial. Trans R Soc Trop Med Hyg 2014;**108**(1):13-21 doi: 10.1093/trstmh/trt103.
70. Totsu S, Yamasaki C, Terahara M, et al. Bifidobacterium and enteral feeding in preterm infants: cluster-randomized trial. Pediatr Int 2014;**56**(5):714-9 doi: 10.1111/ped.12330.
71. Tran KP, Nguyen Q, Truong XN, et al. A comparison of ketamine and morphine analgesia in prehospital trauma care: a cluster randomized clinical trial in rural Quang Tri province, Vietnam. Prehosp Emerg Care 2014;**18**(2):257-64 doi: 10.3109/10903127.2013.851307.
72. Trost SG, Sundal D, Foster GD, et al. Effects of a pediatric weight management program with and without active video games a randomized trial. JAMA Pediatr 2014;**168**(5):407-13 doi: 10.1001/jamapediatrics.2013.3436.
73. Umanodan R, Shimazu A, Minami M, et al. Evaluation of a Computer-based Stress Management Training Program for Workers' Psychological Well-being and Work Performance: A Cluster Randomized

Controlled Trial. Industrial health 2014

74. Valve P, Lehtinen-Jacks S, Eriksson T, et al. LINDA - a solution-focused low-intensity intervention aimed at improving health behaviors of young females: a cluster-randomized controlled trial. BMC Public Health 2013;**13**:1044 doi: 10.1186/1471-2458-13-1044.
75. van de Steeg L, IJkema R, Langelaan M, et al. Can an e-learning course improve nursing care for older people at risk of delirium: a stepped wedge cluster randomised trial. BMC Geriatr 2014;**14**:69 doi: 10.1186/1471-2318-14-69.
76. Van den Donk M, Griffin SJ, Stellato RK, et al. Effect of early intensive multifactorial therapy compared with routine care on self-reported health status, general well-being, diabetes-specific quality of life and treatment satisfaction in screen-detected type 2 diabetes mellitus patients (ADDITION-Europe): a cluster-randomised trial. Diabetologia 2013 doi: 10.1007/s00125-013-3011-0.
77. Vicens C, Bejarano F, Sempere E, et al. Comparative efficacy of two interventions to discontinue long-term benzodiazepine use: cluster randomised controlled trial in primary care. Br J Psychiatry 2014;**204**(6):471-9 doi: 10.1192/bjp.bp.113.134650.
78. Williams AE, Stevens VJ, Albright CL, et al. The results of a 2-year randomized trial of a worksite weight management intervention. Am J Health Promot 2014;**28**(5):336-9 doi: 10.4278/ajhp.100127-ARB-29.
79. Williams SE, Rothman RL, Offit PA, et al. A randomized trial to increase acceptance of childhood vaccines by vaccine-hesitant parents: a pilot study. Acad Pediatr 2013;**13**(5):475-80 doi: 10.1016/j.acap.2013.03.011.
80. Wilson A, O'Hare JP, Hardy A, et al. Evaluation of the clinical and cost effectiveness of intermediate care clinics for diabetes (ICCD): a multicentre cluster randomised controlled trial. PLoS One 2014;**9**(4):e93964 doi: 10.1371/journal.pone.0093964.
81. Wilson GB, Wray C, McGovern R, et al. Intervention to reduce excessive alcohol consumption and improve comorbidity outcomes in hypertensive or depressed primary care patients: two parallel cluster randomized feasibility trials. Trials 2014;**15**:235 doi: 10.1186/1745-6215-15-235.
82. Wolfenden L, Wyse R, Campbell E, et al. Randomized controlled trial of a telephone-based intervention for child fruit and vegetable intake: long-term follow-up. Am J Clin Nutr 2014;**99**(3):543-50 doi: 10.3945/ajcn.113.071738.
83. Wüsthoff LE, Waal H, Gråwe RW. The effectiveness of integrated treatment in patients with substance use disorders co-occurring with anxiety and/or depression--a group randomized trial. BMC Psychiatry 2014;**14**:67 doi: 10.1186/1471-244X-14-67.
84. Zatzick D, Donovan DM, Jurkovich G, et al. Disseminating alcohol screening and brief intervention at trauma centers: a policy-relevant cluster randomized effectiveness trial. Addiction 2014;**109**(5):754-65 doi: 10.1111/add.12492.
85. Zheng Y, Li XG, Wang QZ, et al. Enhancement of vitamin A combined vitamin D supplementation on immune response to Bacille Calmette-Guérin vaccine revaccinated in Chinese infants. Asian Pac J Trop Med 2014;**7**(2):130-5 doi: 10.1016/S1995-7645(14)60008-0.
86. Zlotkin S, Newton S, Aimone AM, et al. Effect of iron fortification on malaria incidence in infants and young children in Ghana: a randomized trial. JAMA 2013;**310**(9):938-47 doi: 10.1001/jama.2013.277129.

REFERENCES

[1] Marion K Campbell, Jill Mollison, Nick Steen, Jeremy M Grimshaw, and Martin Eccles. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice*, 17(2):192–196, 2000.

[2] Allan Donner and Neil Klar. *Design and analysis of cluster randomization trials in health research*. John Wiley & Sons, 2000.

[3] Marion K Campbell and Jeremy M Grimshaw. Cluster randomised trials: time for improvement. *BMJ-British Medical Journal-International Edition*, 317(7167):1171–1171, 1998.

[4] Richard Hayes and L Moulton. *Cluster randomised trials*. Chapman & Hall/CRC, 2009.

[5] Harold C Sox and Steven N Goodman. The methods of comparative effectiveness research. *Annual Review of Public Health*, 33:425–445, 2012.

[6] Sandra Eldridge and Sally Kerry. *A practical guide to cluster randomised trials in health services research*, volume 120. John Wiley & Sons, 2012.

[7] Allan Donner and Neil Klar. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94(3):416–422, 2004.

[8] Robert L Wears. Advanced statistics: Statistical methods for analyzing cluster and cluster-randomized data. *Academic emergency medicine*, 9(4):330–341, 2002.

[9] David M Murray. *Design and analysis of group-randomized trials*, volume 29. Oxford University Press, USA, 1998.

[10] Jerome Cornfield. Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108(2):100–102, 1978.

[11] Sally M Kerry and J Martin Bland. Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Statistics in Medicine*, 20(3):377–390, 2001.

[12] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

[13] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012.

[14] Jinhui Ma, Parminder Raina, Joseph Beyene, and Lehana Thabane. Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized

trials with missing binary outcomes: a simulation study. *BMC Medical Research Methodology*, 13(1):1, 2013.

[15] MJ Campbell, A Donner, and N Klar. Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26(1):2–19, 2007.

[16] Marion K Campbell, Diana R Elbourne, and Douglas G Altman. Consort statement: extension to cluster randomised trials. *BMJ*, 328(7441):702–708, 2004.

[17] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press, Washington DC, 2010.

[18] Sally Hollis and Fiona Campbell. What is meant by intention to treat analysis? survey of published randomised controlled trials. *BMJ*, 319(7211):670–674, 1999.

[19] Angela M Wood, Ian R White, and Simon G Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376, 2004.

[20] Jocelyn Gravel, Lucie Opatrny, and Stan Shapiro. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? *Clinical Trials*, 4(4):350–356, 2007.

[21] Shona Fielding, Graeme Maclennan, Jonathan A Cook, and Craig R Ramsay. A review of rcts in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials*, 9(51):6215–9, 2008.

[22] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in rcts; a review of the top medical journals. *BMC Medical Research Methodology*, 14(1):1, 2014.

[23] James D Dziura, Lori A Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med*, 86(3):343–358, 2013.

[24] Kenneth F Schulz and David A Grimes. Sample size slippages in randomised trials: exclusions and the lost and wayward. *The Lancet*, 359(9308):781–785, 2002.

[25] Diane L Fairclough. *Design and analysis of quality of life studies in clinical trials*. CRC press, 2010.

[26] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[27] Melanie L Bell and Diane L Fairclough. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Statistical Methods in Medical Research*, page 0962280213476378, 2013.

[28] Michael G Kenward and Geert Molenberghs. Last observation carried forward: a crystal ball? *Journal of Biopharmaceutical Statistics*, 19(5):872–888, 2009.

[29] James R Carpenter and Michael G Kenward. Missing data in randomised controlled trials-a practical guide. *London School of Hygiene*, 2007.

[30] Monica Taljaard, Allan Donner, and Neil Klar. Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, 50(3):329–345, 2008.

[31] Stef Van Buuren et al. *Multiple imputation of multilevel data*. Routledge New York, NY, 2011.

[32] Rebecca R Andridge. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal*, 53(1):57–74, 2011.

[33] Jinhui Ma, Noori Akhtar-Danesh, Lisa Dolovich, and Lehana Thabane. Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Medical Research Methodology*, 11(1):1, 2011.

[34] Jinhui Ma, P Raina, J Beyene, and L Thabane. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *J Open Access Med Stat*, 2:93–103, 2012.

[35] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[36] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[37] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013.

[38] Ian R White, Nicholas J Horton, James Carpenter, Stuart J Pocock, et al. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ*, 342:d40, 2011.

[39] Mallorie H. Fiero, Shuang Huang, Eyal Oren, and Melanie L. Bell. Statistical analysis and handling of missing data in cluster randomised trials: a systematic review. *Trials*, 17:72, 2015.

[40] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[41] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[42] Roderick JA Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

[43] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

[44] Herbert Thijs, Geert Molenberghs, Bart Michiels, Geert Verbeke, and Desmond Curran. Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265, 2002.

[45] Hakan Demirtas and Joseph L Schafer. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22(16):2553–2575, 2003.

[46] Ali Satty and Henry Mwambi. Selection and pattern mixture models for modelling longitudinal data with dropout: An application study. *SORT: Statistics and Operations Research Transactions*, 37(2):131–152, 2013.

[47] Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.

[48] Geert Molenberghs, Bart Michiels, Michael G Kenward, and Peter J Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161, 1998.

[49] Ofer Harel and Joseph L Schafer. Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50, 2009.

[50] Wendy J Post, Ciska Buijs, Ronald P Stolk, Elisabeth GE de Vries, and Saskia Le Cessie. The analysis of longitudinal quality of life measures with informative dropout: a pattern mixture approach. *Quality of Life Research*, 19(1):137–148, 2010.

[51] Rebecca R Andridge and Roderick JA Little. Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27(2):153, 2011.

[52] Allan Donner. Some aspects of the design and analysis of cluster randomization trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1):95–113, 1998.

[53] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[54] Judy M Simpson, Neil Klar, and A Donnor. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, 85(10):1378–1383, 1995.

[55] Karla Díaz-Ordaz, Michael G Kenward, Abie Cohen, Claire L Coleman, and Sandra Eldridge. Are missing data adequately handled in cluster randomised trials? a systematic review and guidelines. *Clinical Trials*, page 1740774514537136, 2014.

[56] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of Internal Medicine*, 151(4):264–269, 2009.

[57] Mallorie Fiero, Shuang Huang, and Melanie L Bell. Statistical analysis and handling of missing data in cluster randomised trials: protocol for a systematic review. *BMJ Open*, 5(5):e007378, 2015.

[58] Michael A Hussey and James P Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2):182–191, 2007.

[59] Karla Hemming, TP Haines, PJ Chilton, AJ Girling, and RJ Lilford. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350:h391, 2015.

[60] Anthony Shakeshaft, Christopher Doran, Dennis Petrie, Courtney Breen, Alys Havard, Ansari Abudeen, Elissa Harwood, Anton Clifford, Catherine D'Este, Stuart Gilmour, et al. The effectiveness of community action in reducing risky alcohol consumption and harm: a cluster randomised controlled trial. *PLoS Med*, 11(3):e1001617, 2014.

[61] Lisa M Sullivan, Kimberly A Dukes, and Elena Losina. Tutorial in biostatistics. an introduction to hierarchical linear modelling. *Statistics in Medicine*, 18(7):855–888, 1999.

[62] Neil W Scott, Gladys C McPherson, Craig R Ramsay, and Marion K Campbell. The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials*, 23(6):662–674, 2002.

[63] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 1987.

[64] Juned Siddique, Ofer Harel, and Catherine M Crespi. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *The Annals of Applied Statistics*, 6(4):1814, 2012.

[65] S Van Buuren and CGM Oudshoorn. mice: Multivariate imputation by chained equations. r package version 1.16, 2007.

[66] C Jane Morrell, R Warner, P Slade, S Dixon, S Walters, G Paley, and T Brugha. *Psychological interventions for postnatal depression: cluster randomised trial and economic evaluation: the PoNDER trial.* Prepress Projects, 2009.

[67] Stef Van Buuren, Hendriek C Boshuizen, Dick L Knook, et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999.

[68] Ian R White, James Carpenter, Stephen Evans, and Sara Schroter. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials*, 4(2):125–139, 2007.

[69] Geoffrey Adams, Martin C Gulliford, Obioha C Ukoumunne, Sandra Eldridge, Susan Chinn, and Michael J Campbell. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57(8):785–794, 2004.

[70] Nicole Morgan Gibson and Stephen Olejnik. Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, 63(2):204–238, 2003.

[71] Mike W-L Cheung. Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random. *Organizational Research Methods*, 2007.

[72] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[73] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[74] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011.