Quantifying Power and Bias in Cluster Randomized Trials Using Mixed Models vs Cluster-level Analysis in the Presence of Missing Data: A Simulation Study

Brenda M Vincent[1,2] and Melanie L Bell[1]
[1] Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ
[2] VA Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, MI

Abstract

Background/Aims: The two main approaches used to analyze cluster randomized trials are cluster-level and individual-level analysis. In a cluster-level analysis, summary measures are obtained for each cluster and then the two sets of cluster-specific measures are compared, such as with a t-test of the cluster means. A mixed model which takes into account cluster membership is an example of an individual-level analysis. The purpose of this study was to compare power and bias of a cluster-level analysis and an individual level analysis when data are complete, missing completely at random (MCAR) and missing at random (MAR).

Methods: We used a simulation study to quantify and compare power and bias of these two methods. Complete datasets were generated and then data were deleted to simulate MCAR and MAR data. A balanced design, with two treatment groups and two time points was assumed. Cluster size, variance components (including within-subject, within-cluster and between-cluster variance) and proportion missing were varied to simulate common scenarios seen in practice. For each combination of parameters, 1000 datasets were generated and analyzed.

Results: Results of our simulation study indicate that cluster-level analysis resulted in substantial loss of power (up to 26%) when data were MAR. Individual-level analysis had higher power and remained unbiased, even with a small number of clusters.

Conclusion: Individual-level modeling which takes into account cluster membership performs better in the presence of missing data in terms of power and bias.

Key Words: cluster randomized trial, power, bias, missing data, mixed model

1. Introduction

In cluster randomized trials (CRTs), intact groups of individuals are allocated to treatment arms while the outcome of interest is assessed on the individuals. A CRT may be adopted to reduce contamination that may occur if individuals in the same community are assigned to different treatment arms; if an intervention is given to an entire group

either by design or for logistical convenience; or to assess the population-level effects of an intervention applied to a large proportion of a population. [1]

Special consideration must be taken in the analysis of CRTs because the unit of randomization (clusters) can be different than the level at which data are collected (individuals). Observations on individuals within clusters tend to be correlated because individuals within clusters tend to be more similar than in a randomly selected sample. Ignoring the dependence during analysis can lead to underestimated standard errors. [2] Since subjects within clusters cannot be treated as independent, analytical approaches which take into account the cluster design are necessary.[3] The intracluster correlation coefficient (ICC) is used to measure how much more similar observations within a cluster are compared to observations between clusters. The ICC is defined as the proportion of total variance that can be attributed to the differences between clusters. [4] This correlation is generally small in CRTs, typically ranging from 0.001 and 0.05.[5, 6] However, even small ICCs may have a large impact on the power of a study due to the reduction in effective sample size.[7]

1.1 Approaches for analyzing CRTs

There are two main approaches that are generally used to analyze CRTs: cluster-level and individual-level analysis. Individual-level analyses generally use a regression model that takes into account the correlation of individuals within clusters, such as linear mixed models or generalized estimating equations.[6, 8] Cluster membership and a subject indicator (if data are longitudinal) are included as random effects in a mixed model approach. [1] Some strengths of individual level modeling include the ability to use individual level covariates in the model as well as modelling the variance between-cluster and within-cluster, which may lead to a more realistic model of the clustered data.[6, 9] Individual-level modelling tends to be more efficient when cluster size varies substantially.[1, 10] Even in a design in which equal-sized clusters are assigned, cluster sizes may vary substantially in the presence of missing data.

With cluster-level analysis, summary measures are obtained for each cluster and then the cluster-specific measures are compared.[11] For continuous outcomes, the means of each cluster can be compared using a t-test. Strengths of cluster-level analysis are the relative simplicity of the analysis and its capacity for handling small numbers of clusters; cluster-level methods have been recommended when there are fewer than 15-20 clusters per treatment arm because they are more robust to departures from underlying assumptions.[1] It has been shown that cluster-level analyses are robust with as few as three clusters when using a t-test.[4] However, a cluster-level analysis has a small number of degrees of freedom (the number of clusters minus two) leading to low statistical efficiency,[11] which may be heightened in the presence of missing data.

1.2 Missing data

Missing data are common in both individually randomized designs [12] and CRTs [13] and can be classified as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR).[14] Under MCAR, the probability that a response is missing does not depend on observed or unobserved data. With MAR data, the probability that a response is missing may depend on observed data, but not on unobserved data. If the probability of a missing response is related to unobserved data, the missingness is termed MNAR.[14] The potential consequences of missing data are a loss of power and biased estimation. Under MCAR, an analysis on the observed outcomes only (referred to as complete case analysis) results in unbiased estimates.[15]

When data are MAR, mixed models that adjust for the covariates that are associated with the missing data mechanism yield unbiased estimates.[14, 15]

Taljaard et al. compared the Type I and Type II error rates of various imputation techniques for handling missing data in CRTs using simulation studies,[16] but they did not compare individual versus cluster level analysis. Hossain et al compared the performance of cluster-level analysis, baseline covariate-adjusted cluster-level analysis and a mixed model under covariate-dependent missing data for bias, average standard errors and coverage.[17] However, they did not consider other missing data mechanisms including MCAR and MAR data. Furthermore, they did not compare the power between individual-level and cluster-level analysis.

1.3 Objective

To our knowledge, there has not been a comparison of cluster-level and individual level analysis in the presence of MCAR or MAR missing data. We carried out a simulation study to quantify power and bias in CRTs when using individual-level analysis compared to cluster-level analysis with complete data and in the presence of MCAR and MAR missing data. We investigated the effect of varying intracluster correlation, cluster size, and the proportion of missingness.

2. Methods

We simulated data from a balanced CRT with two treatment groups and two time points. We used a two-sided test, $\alpha = 0.05$, power $\varphi = 0.8$, total variance = 4, and 20 subjects per cluster in the initial power analysis to design the study. See Appendix for more details. The treatment effect was varied to maintain 80% power. The following combinations were used: number of clusters = $20, 60$; ICC = $0.001, 0.01, 0.05$; and proportion missing = $0.2, 0.4$. For each combination, a complete dataset was simulated and analyzed using both a mixed model and a cluster-level t-test. Observations were then removed to simulate MCAR and MAR data and reanalyzed. For each scenario 1,000 datasets were simulated and analyzed. Bias and power for the treatment effect estimate were assessed. Details are given below.

2.1 Data Simulation

For each individual, the outcome Y_{ijk} was simulated using the model:[18]

$$Y_{ijk} = \beta_0 + \tau T_k + \delta X_i T_k + c_i + s_{j(i)} + e_{ijk}$$

where $i = 1, 2, \ldots, n$ is the index for the cluster, $j = 1, 2, \ldots, m$ is the index for subjects nested within each cluster, and $k = 0, 1$ is the index for time. The treatment indicator is X_i, and the time indicator is T_k. For the purposes of this study, $\beta_0 = 5$, $\tau = 0.5$, and the treatment effect, δ, varied in order to target 80% power in the complete data, as described above. Clusters, c_i were generated from a normal distribution with mean 0 and variance σ_C^2. Within each cluster, subjects, $s_{j(i)}$, were sampled from a normal distribution with mean 0 and variance σ_s^2. Each subject had a normally distributed error, e_{ijk}, with mean 0 and σ_e^2. The variance components were varied to create the different ICCs, while keeping the total variance at 4. The between subject variance, σ_s^2, was kept constant at 3 while σ_C^2 and σ_e^2 were varied such that:

$$ICC = \frac{\sigma_C^2}{\sigma_c^2 + \sigma_s^2 + \sigma_e^2} = \frac{\sigma_C^2}{4}$$

Thus, for an ICC of 0.05: $\sigma_C^2 = 0.2, \sigma_s^2 = 3$ and $\sigma_e^2 = 0.8$. For an ICC of 0.01: $\sigma_C^2 = 0.04, \sigma_s^2 = 3$ and $\sigma_e^2 = 0.96$, and for an ICC of 0.01: $\sigma_C^2 = 0.004, \sigma_s^2 = 3$ and $\sigma_e^2 = 0.996$.

2.1.1 Simulating Missing Data

The complete data set was the comparison set. Data were deleted to represent MCAR data and MAR data. For the MCAR datasets, the observations at time 1 (follow-up) were randomly deleted throughout with probability of 0.2 or 0.4. For the MAR datasets, the probability of an observation at time 1 being deleted depended on the baseline response value. Specifically, observations with baseline values of greater than 6 were twice as likely to be missing as observations with baseline values of 6 or less. At baseline, the overall mean was 5 with a total variance of 4 for each individual. Thus, 6 is half a standard deviation above the mean. With the proportion of missingness set at 20%, those with baseline values greater than 6 had a 40% probability of having a missing outcome at the next time point and those with baseline values below 6 had a 20% probability of being missing.

2.2 Analysis

The individual-level analysis we used was a mixed model with random effects for cluster membership and subject indicator. Fixed effects included the intervention assignment, time and the interaction between time and intervention group to allow for the effect of the treatment over time to vary by intervention group. For the cluster-level analysis, the mean value was obtained for each cluster and then a two sample t-test with degrees of freedom equal to the total number of clusters minus two was used to compare the cluster means for the two treatment arms.[11]

The estimated treatment effect and associated confidence intervals at the second time point were found for each type of analysis. Power and bias were assessed and compared between the individual-level analysis method and the cluster-level analysis method for the three types of datasets (complete, MCAR and MAR). We calculated power as the proportion of statistically significant results, defined as $p<0.05$. We calculated bias as the difference between the estimated treatment effect from the true treatment effect. Both absolute bias and percent bias were calculated.

3. Results

For each combination of parameters (ICC, number of clusters and probability of missingness), 1000 datasets were simulated. Power and bias were calculated. Complete datasets were generated and then data were removed to create the MCAR and MAR datasets. Power and bias for the complete datasets were combined and averaged for each set of parameters (because no data were removed), resulting in 2,000 simulation runs for each combination of cluster size and ICC.

3.1 Power

Nominal power was 80% for the complete datasets, and was achieved for each set of parameters (within the 95% confidence intervals). Results are given in the Appendix.

Individual-level analysis was slightly more powerful than cluster-level analysis on the complete datasets; difference in power ranged from 0% to 5%, with a median difference of 1.5%. Under MCAR data, individual-level analysis was more powerful than cluster-level analysis. The difference in power between individual-level and cluster-level analysis ranged from 4 to 16%, with a median difference of 6.5%. With data that were MAR, there was a substantial difference in power between individual-level and cluster-level analysis. Individual level analysis was more powerful by between 2% and 12%, with a median difference of 7.5%.

Figures 1-3 show how power was impacted by ICC, number of clusters and proportion of missingness when data were MCAR and MAR. As Figure 1 indicates, individual-level analysis was more powerful overall than cluster level analysis for both MCAR and MAR data. There was a slight increase in power for both individual-level and cluster-level analysis as ICC increased. The loss of power was substantial when data were MAR for cluster-level analysis. Additionally, the range in power was larger for cluster-level analysis, particularly when data were MAR.

In Figure 2, individual and cluster-level power is separated by the total number of clusters. Individual-level analysis was more powerful than cluster-level analysis. There was not a substantial difference in power for 20 clusters compared to 60 clusters for either analysis method when data were MCAR or MAR.

Figure 3 displays power for individual versus cluster-level analysis by the probability that an observation was missing for MCAR and MAR data. There was a substantial loss of power with higher rates of missing data for cluster-level analysis. Individual-level analysis remained more powerful than cluster-level analysis, and was only slightly affected by the increased proportion of missing data.

3.2 Bias

Both individual-level and cluster-level analyses resulted in unbiased estimates when data were complete, MCAR and MAR. The bias was less than 5% for each set of parameters. Figures 4-6 in the Appendix show how ICC, number of clusters and proportion of missingness impact bias when data are MCAR and MAR. Overall, the MCAR datasets were slightly less biased (<2% bias) than the MAR datasets. Results in tabular form are given in the Appendix.

4. Discussion

The objective of our study was to compare individual-level analysis and cluster-level analysis for CRTs on power and bias for complete data and in the presence of missing data. Results from our simulation studies indicate that with complete data, there was not a significant difference in power or bias between the two analysis methods. However, the individual-level mixed model was more powerful than the cluster-level t-test in the presence of missing data. When subjects within clusters were more similar, indicated by increased ICC, there was more power when data were MCAR or MAR. The number of clusters did not have a substantial effect on power for either type of analysis. However, the proportion missing had a large impact on power for cluster-level analysis, especially when data were MAR. With data that were MCAR, there was a similar trend in the loss

of power that resulted from having a decreased sample size for both individual-level and cluster-level analysis.

With the complete datasets, MCAR datasets and MAR datasets, both analysis methods resulted in unbiased estimates. The ICC and the number of clusters did not have a significant effect on bias under either missingness assumption.

The results of our simulations are supported by the literature in terms of power. Ashbeck and Bell [19] demonstrated that mixed models resulted in greater power than t-tests in the presence of missing data for individually randomized studies. It has been shown that mixed models result in unbiased estimates when data are MAR [14, 15, 20] and that a complete case analysis using a t-test for continuous data is valid when data are MCAR. [14, 15, 20] We found that mixed models resulted in unbiased estimates for MAR data; however, a complete case analysis was unbiased when data were MCAR. Hossain et al [17] also found that cluster-level analysis and mixed models resulted in unbiased estimates when the missingness depended on covariates. Hossain et al [17] showed that a cluster-level analysis was unbiased if both treatment groups have the same missingness mechanisms and covariate effects, which is the scenario our simulations fell in.

Researchers using a CRT design can use the results of this study to minimize the loss of power and bias that may occur due to missing data. The risk of attrition is high in CRTs, [13, 21] and thus it is almost inevitable that missing data will exist. In practice, it is difficult to determine which missingness assumption is appropriate. [22] However, it is important to plan for missing data in the design and analysis stages of CRTs. There is a trade-off that should be considered when deciding which analysis technique to adopt. Cluster-level analysis is relatively simple and tends to do well in terms of power and bias if data are MCAR. However, when data are MAR, cluster-level analysis resulted in a substantial loss of power. Researchers can consider which scenario fits their study in terms of ICC and missingness mechanism and adopt the analysis method that is most appropriate. The use of a mixed model, though more complex and difficult to employ, may be advantageous because it was more efficient regardless of the missingness mechanism.

A limitation of our study, like all simulations studies, is that it was relatively narrow in scope. We considered a limited number of factors which may have an impact on power and bias. Given the parameters of our study, we had surprising results. Cluster-level methods are recommended when there are fewer than 15-20 clusters per treatment arm because cluster-level analyses are more robust to departures from underlying assumptions. [1] The results of our study indicate that individual-level analysis performed better in terms of power, even with 10 clusters per arm. In our simulations, we used mixed models and one of the assumptions is that random effects are normally distributed. Through our simulations, we did not misspecify the distribution so this assumption was always met, which may not reflect reality. Because the assumption of normally distributed random effects was met with our simulations, this may partially explain why cluster-level analysis remained less efficient than individual-level analysis, even with a smaller number of clusters per arm. Further investigation is needed to assess whether the recommendation to use cluster level analysis with a smaller number of clusters is appropriate.

Strengths of our study included use of simulations to quantify the difference in power between the two analysis methods. An intuitive understanding seems to be given in the literature that cluster-level analysis is less efficient due to the decreased sample size and

underestimated variance; however an in-depth investigation has not yet been provided. Additionally, we focused on the effect of several factors that may influence the power and bias.

In this study, we used the simple case of a balanced CRT design to compare individual-level analysis to cluster-level analysis. We were then able to systematically vary the components of interest including the missingness mechanism, ICC, number of clusters and proportion missing. Future investigation may be warranted for the case of CRTs with varying cluster sizes and for categorical outcomes. To summarize, we found that individual-level analysis was more efficient than cluster-level analysis and resulted in unbiased estimates. If possible, researchers should use individual level analyses.

References

1. Hayes RJ and Moulton LH. *Cluster randomised trials*. Boca Raton [etc.]: Boca Raton [etc.], 2009.
2. Chuang J, Hripcsak G and Heitjan DF. Design and Analysis of Controlled Trials in Naturally Clustered Environments: Implications for Medical Informatics 2002; 9: 230-238.
3. Bland JM and Kerry SM. Statistics notes. Trials randomised in clusters 1997; 315: 600-600.
4. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials 2002; 21.
5. Donner A. Some aspects of the design and analysis of cluster randomization trials 1998; 47: 95-113.
6. Fitzmaurice GM, Laird NM and Ware JH. *Applied Longitudinal Analysis*. Hoboken NJ: Hoboken NJ, 2011.
7. Klar DA. *Design and analysis of cluster randomization trials in health research*, 2000.
8. Rotnitzky A and Jewell N. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 1990; 7: 485-497.
9. Zeger SL and Liang K. Longitudinal Data Analysis for Discrete and Continuous Outcomes 1986; 42: 121-130.
10. Hussey M and Hughes J. Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials 2007; 28: 182-191.
11. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data 2002; 9: 330-341.
12. Bell ML, Fiero M, Horton NJ, et al. Handling missing data in RCTs; a review of the top medical journals 2014; 14: 1-8.
13. Fiero MH, Huang S, Oren E, et al. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review 2016; 17: 1-10.
14. Rubin DB. Inference and missing data 1976; 63: 581-592.
15. Bell ML and Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes 2014; 23: 440-459.
16. Taljaard M, Donner A and Klar N. Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials 2008; 50: 329-345.
17. Hossain A, Diaz-Ordaz K and Bartlett JW. Missing continuous outcomes under covariate dependent missingness in cluster randomised trials 2016.
18. Hedeker DR, Gibbons RD and Wiley I. *Longitudinal data analysis*. Hoboken, N.J: Hoboken, N.J, 2006.

19. Ashbeck EL and Bell ML. Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data. BMC Medical Research Methodology 2016; 16: 1-8.
20. Council NR, Wartella EA, National Research Council . Panel on Handling Missing Data in Clinical, Trials, et al. *The prevention and treatment of missing data in clinical trials*. Washington, D.C: Washington, D.C, 2010.
21. Donner A, Brown KS and Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989 1990; 19: 795-800.
22. White IR and Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values 2010; 29: 2920-2931.

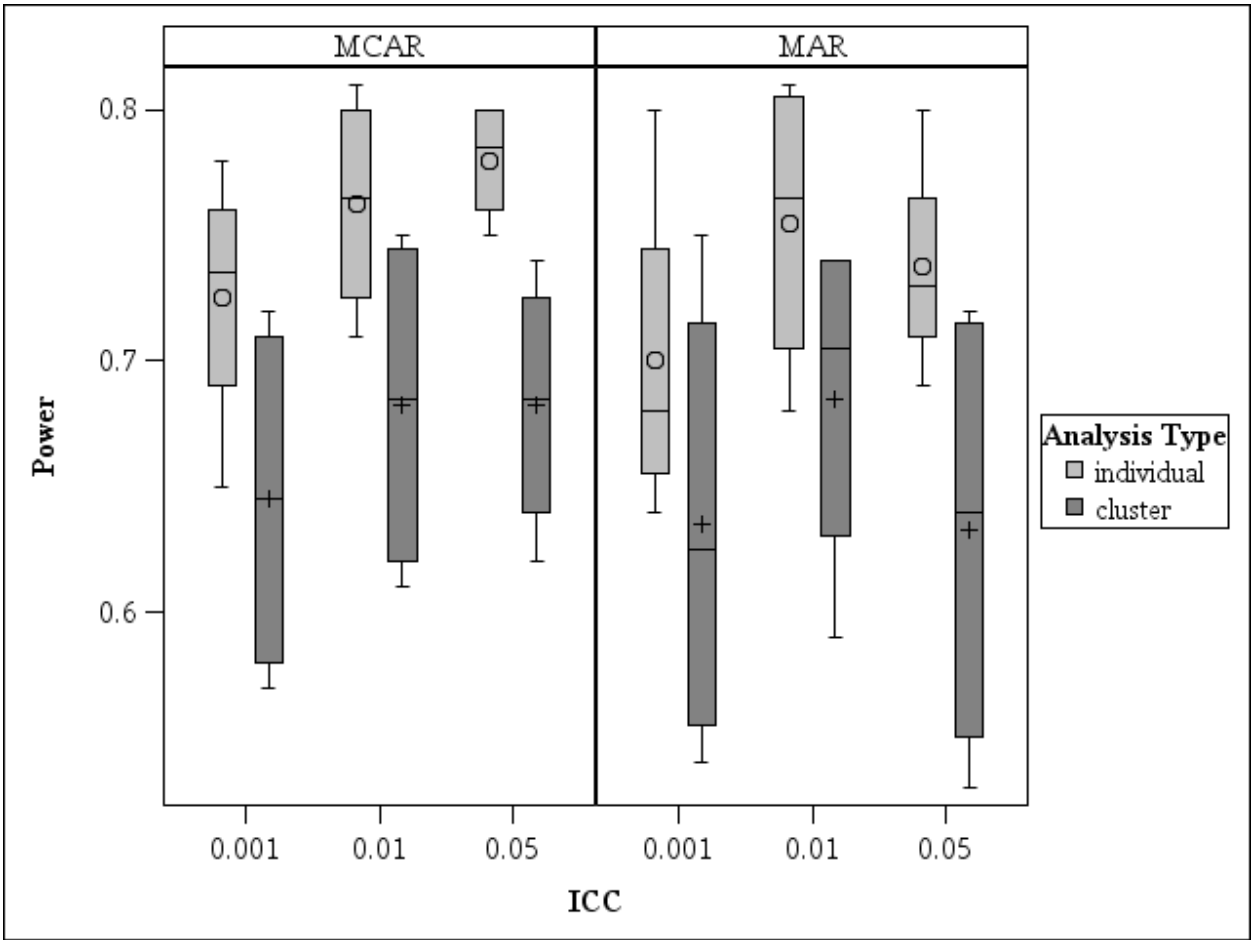

Figure 1: The effect of varying ICC on power. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 4,000 simulation trials with number of clusters (20, 60) and probability of missing (20%, 40%). Nominal power was 80% for complete data. The boxes show min, max, first and third quartile, median (line) and mean (symbol).

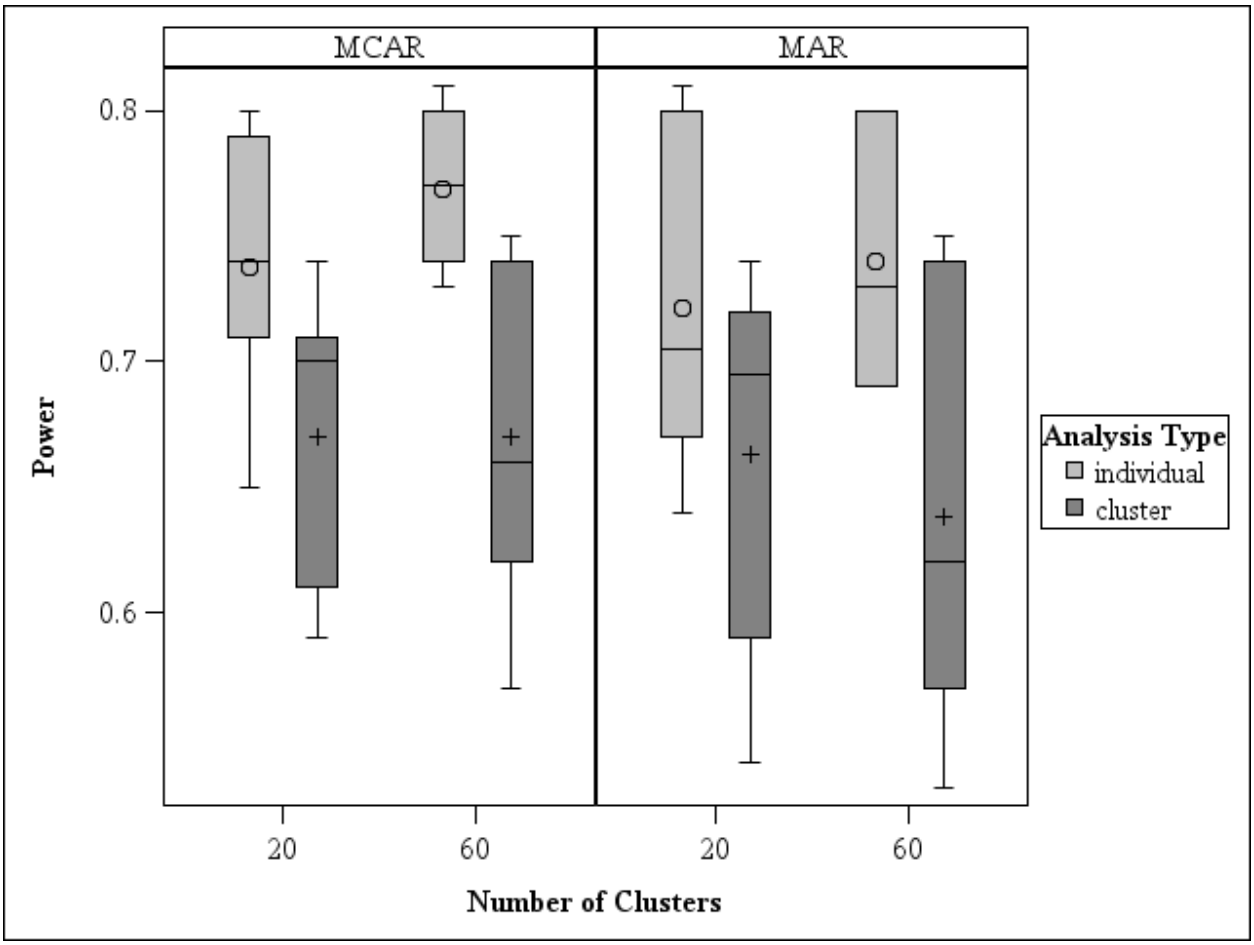

Figure 2: The effect of varying number of clusters on power. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 6,000 simulation trials with ICC (.001, .01, and .05) and probability of missing (20%, 40%). Nominal power was 80% for complete data. The boxes show min, max, first and third quartile, median (line) and mean (symbol).

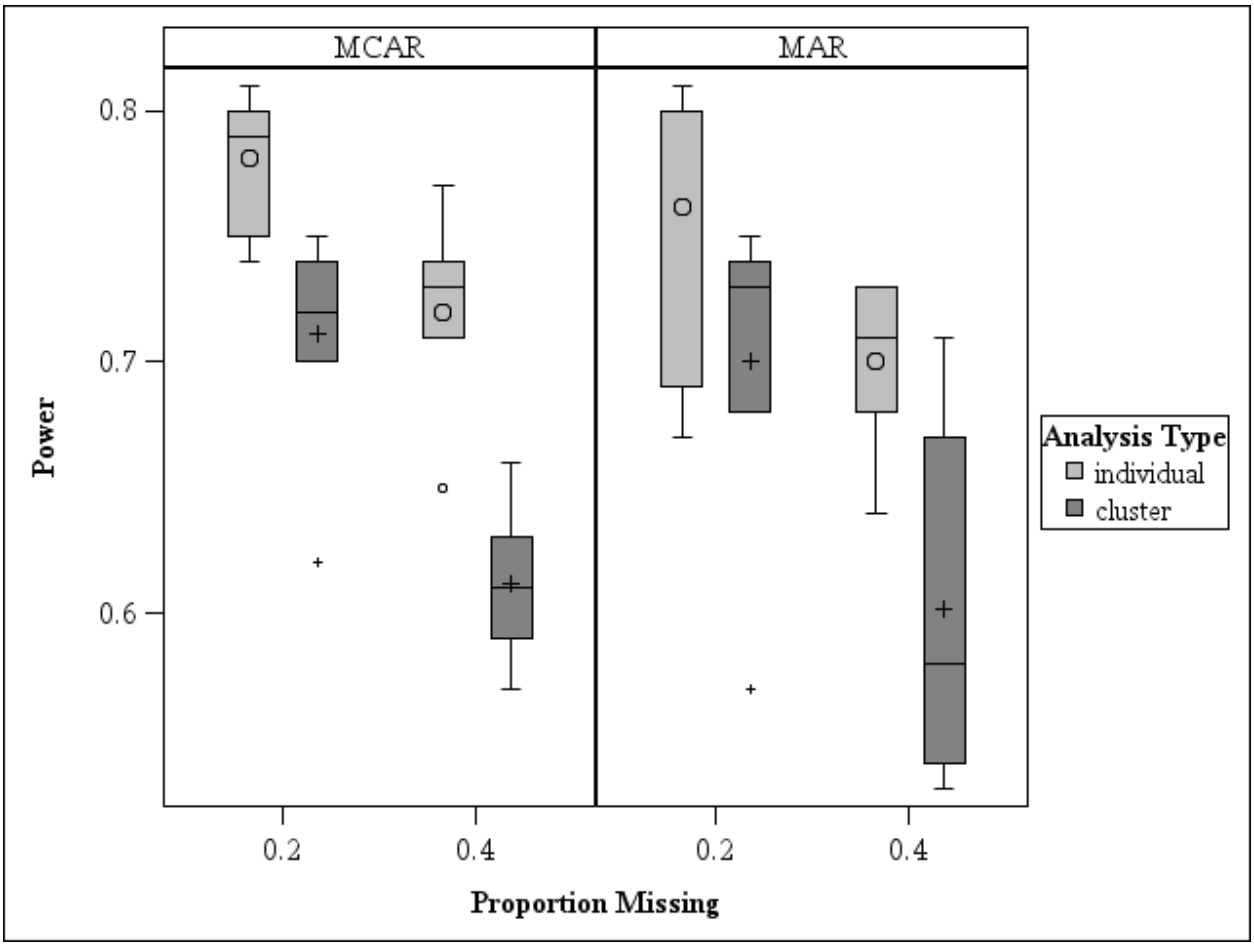

Figure 3: The effect of varying proportion missing on power. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 6,000 simulation trials with ICC (0.001, 0.01, and 0.05) and number of clusters (20 and 60). Nominal power was 80% for complete data. The boxes show min, max, first and third quartile, median (line) and mean (symbol).

Appendix

Table 1: Power and bias for the treatment effect. Results from 2,000 simulation trials in the complete datasets.

ICC	Clusters	Power (95% Confidence Interval)		Absolute Bias (Percent Bias)	
		Individual-Level	Cluster-Level	Individual-Level	Cluster-Level
.001	20	0.81(0.80, 0.83)	0.80(0.78,0.81)	0.00(0.2)	0.00(0.1)
.01		0.81(0.79, 0.83)	0.80(0.78,0.81)	-0.01(-1.1)	0.00(0.1)
.05		0.81(0.80, 0.83)	0.76(0.74,0.78)	0.00(0.2)	0.00(-0.5)
.001	60	0.82(0.80,0.84)	0.80(0.78,0.82)	0.00(1.1)	0.0(-1.1)
.01		0.82(0.80,0.83)	0.82(0.80,0.83)	0.00(-0.8)	0.00(0.6)
.05		0.80(0.79,0.82)	0.83(0.82,0.85)	0.00(0.5)	0.00(-0.1)

ICC, number of clusters, and proportion of missingness* were varied and 1,000 simulation runs were performed.
Power was defined as the proportion of trials where $p<0.05$
Bias is the difference between the treatment effect and the average estimate.
*Because no data are removed for the complete datasets, the 20% and 40% missingness were combined and the average power and bias were taken for 2,000 datasets

Table 2: Power and bias for the treatment effect. Results from 1,000 simulation trials in the MCAR datasets.

ICC	Clusters	Probability of missing	Power (95% Confidence Interval)		Absolute Bias (Percent Bias)	
			Individual-Level	Cluster-Level	Individual-Level	Cluster-Level
.001	20	0.20	0.74(0.72,0.77)	0.70(0.68,0.73)	0.01(1.6)	0.00(-0.6)
.01			0.79(0.76,0.81)	0.74(0.71,0.76)	0.00(-0.1)	-0.01(-0.9)
.05			0.80(0.78,0.83)	0.71(0.68,0.74)	0.00(-0.6)	0.00(-0.1)
.001	20	0.40	0.65(0.62,0.68)	0.59(0.56,0.62)	0.01(1.0)	0.00(0.5)
.01			0.71(0.69,0.74)	0.61(0.57,0.64)	0.00(-0.5)	0.00(0.1)
.05			0.75(0.72,0.78)	0.62(0.59,0.65)	0.00(-0.1)	-0.01(-1.1)
.001	60	0.20	0.78(0.75,0.80)	0.72(0.70,0.75)	0.00(1.4)	0.00(0.8)
.01			0.81(0.79,0.84)	0.75(0.73,0.78)	0.00(0.3)	0.00(0.8)
.05			0.80(0.77,0.82)	0.74(0.72,0.77)	0.00(-0.6)	0.00(0.6)
.001	60	0.40	0.73(0.71,0.76)	0.57(0.54,0.60)	0.00(1.3)	0.01(1.7)
.01			0.74(0.71,0.77)	0.63(0.60,0.66)	0.00(-0.1)	0.00(0.0)
.05			0.77(0.74,0.80)	0.66(0.63,0.69)	0.00(0.9)	0.00(-0.1)

ICC, number of clusters, and proportion of missingness were varied and 1,000 simulation runs were performed.
Power was defined as the proportion of trials where $p<0.05$
Bias is the difference between the treatment effect and the average estimate.

Table 3: Power and bias for the treatment effect. Results from 1,000 simulation trials in the MAR datasets.

ICC	Clusters	Probability of missing	Power (95% Confidence Interval)		Absolute Bias (Percent Bias)	
			Individual-Level	Cluster-Level	Individual-Level	Cluster-Level
.001	20	0.20	0.70(0.67,0.73)	0.68(0.65,0.71)	-0.02(-2.7)	0.00(0.2)
.01			0.81(0.78,0.83)	0.74(0.71,0.76)	0.03(4.6)	-0.02(-2.5)
.05			0.80(0.77,0.82)	0.72(0.69,0.75)	0.01(0.8)	-0.01(-1.0)
.001	20	0.40	0.64(0.61,0.67)	0.54(0.51,0.57)	-0.01(-3.5)	0.00(0.4)
.01			0.68(0.65,0.70)	0.59(0.56,0.62)	0.00(0.4)	0.00(0.0)
.05			0.73(0.70,0.76)	0.63(0.60,0.66)	0.00(0.6)	0.00(0.0)
.001	60	0.20	0.73(0.70,0.76)	0.71(0.68,0.73)	0.00(-0.3)	-0.01(-2.1)
.01			0.80(0.78,0.83)	0.75(0.72,0.77)	-0.01(-1.1)	-0.01(-1.8)
.05			0.80(0.78,0.83)	0.74(0.71,0.76)	-0.01(-0.9)	0.00(0.5)
.001	60	0.40	0.69(0.66,0.72)	0.57(0.64,0.60)	0.01(1.8)	0.00(-1.3)
.01			0.69(0.66,0.72)	0.57(0.64,0.60)	0.00(-0.6)	0.00(1.0)
.05			0.73(0.70,0.75)	0.67(0.64,0.70)	-0.01(-1.1)	0.00(1.0)

ICC, number of clusters, and proportion of missingness were varied and 1,000 simulation runs were performed.
Power was defined as the proportion of trials where $p<0.05$
Bias is the difference between the treatment effect and the average estimate.

Power Calculation

For each combination of sample size and ICC, a treatment effect was simulated for 80% nominal power to detect a difference between the treatment groups. The total sample size was divided by the design effect (for clusters of size m, DE=1+ (m-1)ICC) [11] , which represents the total sample size needed under a randomized clinical trial with the individual as the unit of randomization and analysis. The individual-design sample size was used to determine the required treatment effect. For example, given 20 clusters with a cluster size of 20 and an ICC of .05, the design effect is $1 + (\mathrm{m} - 1)\mathrm{ICC} = 1 + 19(.05) = 1.95$. This gives an effective sample size of $400/1.95 = 205$. A sample size of 205, α=.05 and standard deviation of 2 (which is $\sqrt{\sigma_c^2 + \sigma_s^2 + \sigma_e^2}$) results in a treatment effect of 0.788 for 80% power.

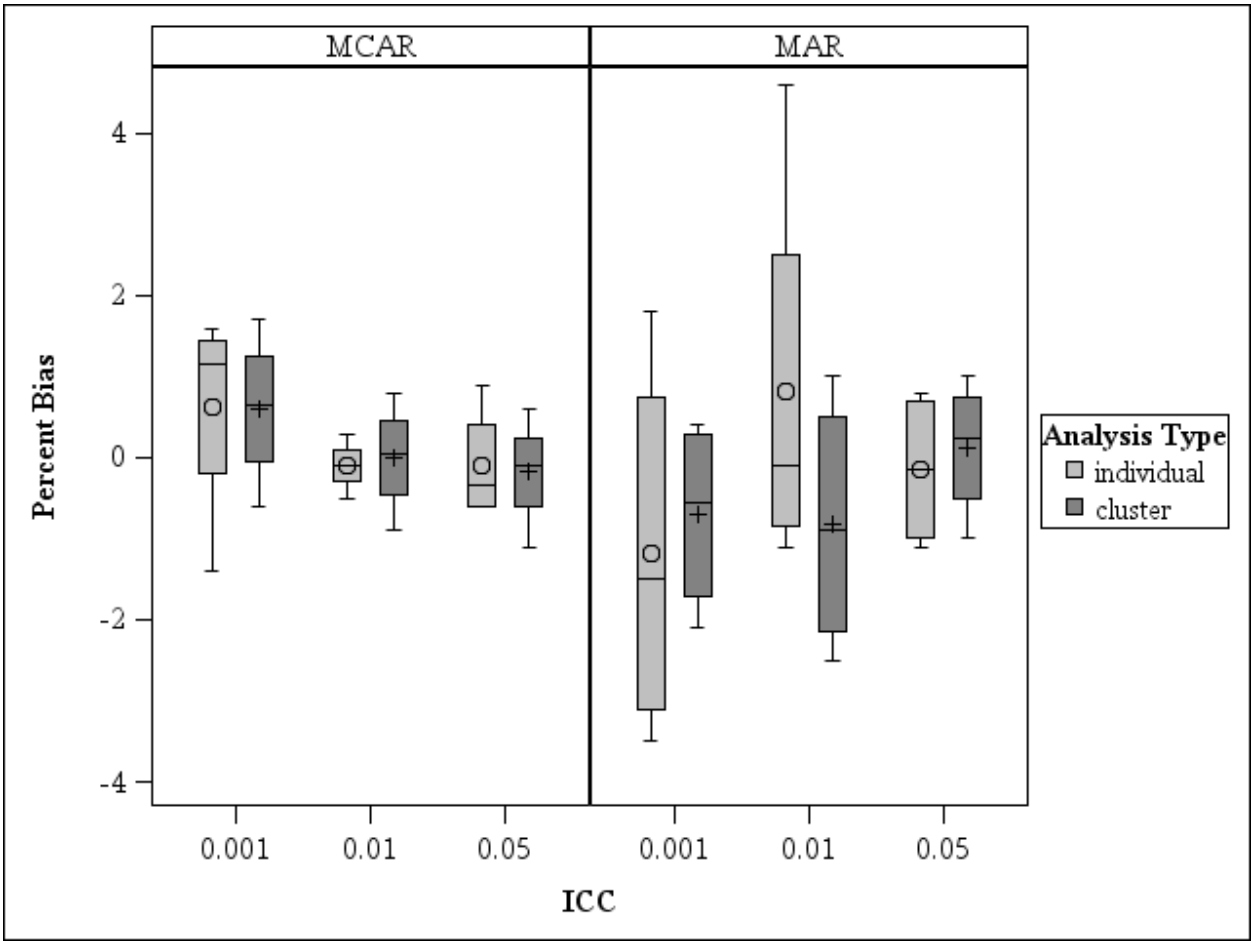

Figure 4: The effect of varying ICC on percent bias. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 4,000 simulation trials with number of clusters (20 and 60) and probability of missing (20% and 40%). The boxes show min, max, first and third quartile, median (line) and mean (symbol).

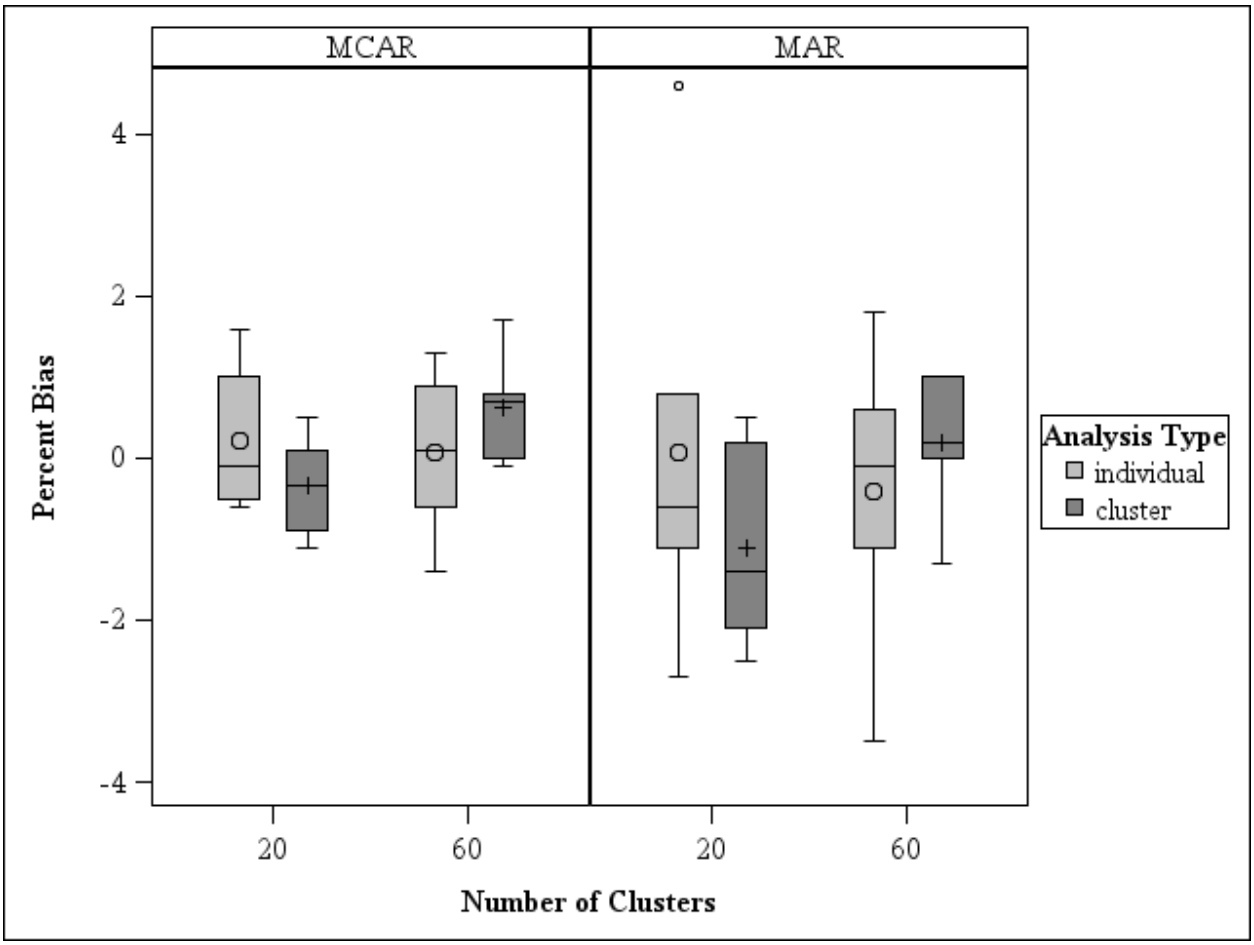

Figure 5: The effect of varying the number of clusters on percent bias. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 6,000 simulation trials with ICC (0.001, 0.01, and 0.05) and probability of missing (20% and 40%). The boxes show min, max, first and third quartile, median (line) and mean (symbol).

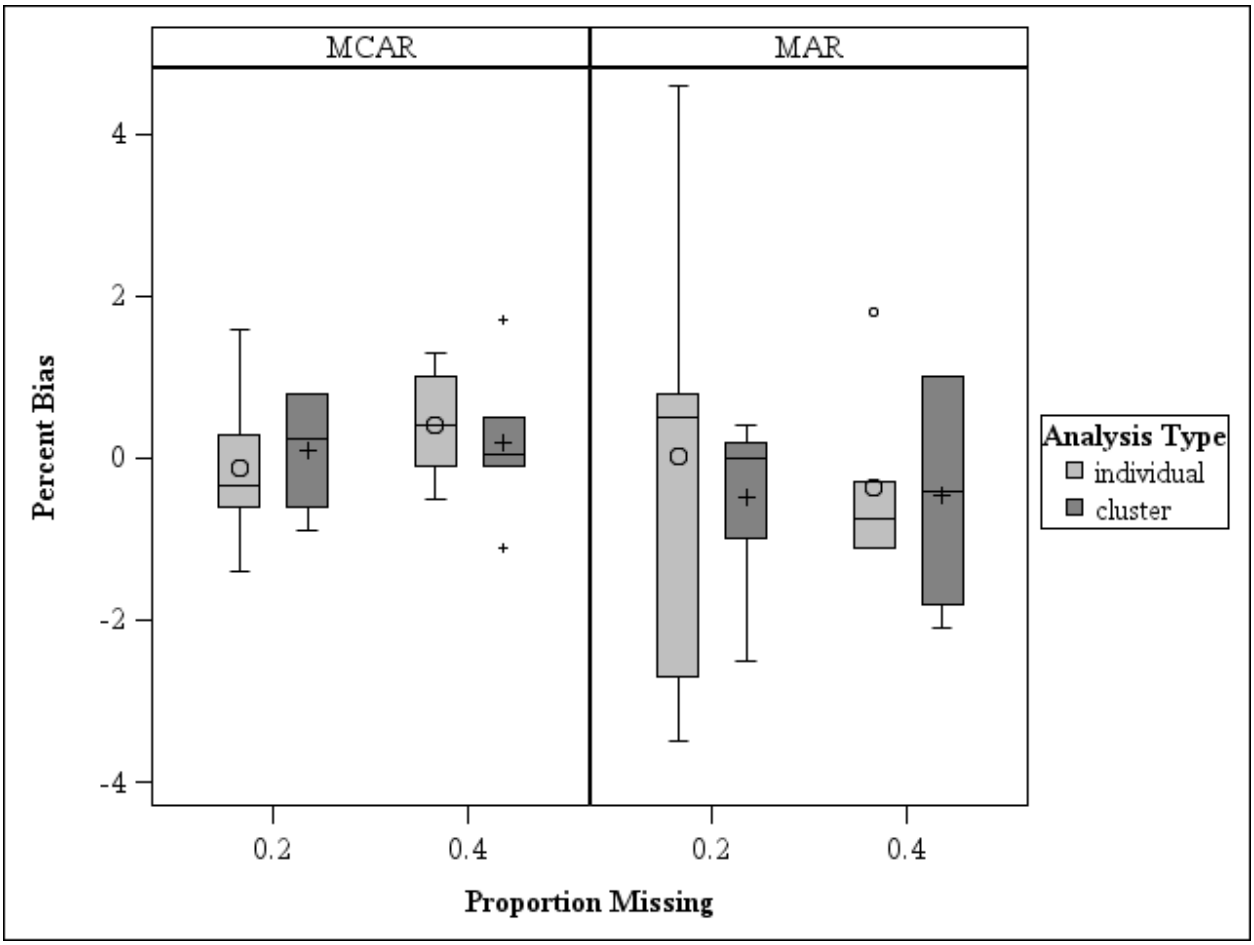

Figure 6: The effect of varying the proportion missing on percent bias. Graphical representation of the combined simulations across the different set of parameters. Each bar represents the power from 6,000 simulation trials with ICC (0.001, 0.01, and 0.05) and number of clusters (20 and 60). The boxes show min, max, first and third quartile, median (line) and mean (symbol).