Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized controlled trials: A Simulation Study

Jinhui Ma, Parminder Raina, Joseph Beyene, Lehana Thabane,

November 11th, 2011

Department of Clinical Epidemiology & Biostatistics
McMaster University
Outline

- Review of missing data strategies for cluster randomized controlled trials (CRTs)
- Strategies focused in this project
- Design of the simulation study
- Results
- Summary of findings
- Limitation
## Missing Data Strategies

### Missing continuous outcomes in CRTs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assumption</strong></td>
<td>Missing completely at random</td>
<td>Missing completely at random and missing at random</td>
</tr>
<tr>
<td><strong>Approach</strong></td>
<td>Simulation</td>
<td>Simulation</td>
</tr>
</tbody>
</table>
| **Strategies Investigated** | 1. Standard regression multiple imputation (MI) ignoring clusters  
2. Cluster mean imputation  
3. Within-cluster MI using approximate Bayesian Bootstrap (ABB) method  
4. Pooled MI using ABB method  
5. Mixed-effects regression MI | 1. Fixed-effects MI incorporating fixed effects for cluster |

## Missing Data Strategies

### Missing binary outcomes in CRTs

<table>
<thead>
<tr>
<th>Study</th>
<th>Ma (2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumption</td>
<td>Missing completely at random and covariate dependent missing</td>
</tr>
<tr>
<td>Approach</td>
<td>Simulation based on a real dataset</td>
</tr>
</tbody>
</table>
| Strategies investigated | 1. Complete case analysis  
2. Standard MI using logistic regression, Markov chain Monte Carlo (MCMC) method, or propensity score method.  
3. Within-cluster MI using logistic regression, MCMC method, or propensity score method.  
4. Across-cluster MI using propensity score method, random-effects logistic regression, or logistic regression with cluster as a fixed effect. |

Objective of This Project

- Investigate the performance of different strategies under different design settings of CRTs and percentage of missing binary outcomes.

- Provide researchers with quantitative evidence to guide the selection of appropriate missing data strategies.
Investigated strategies

- Complete case analysis
  - Only subjects with completed data are included for analysis while subjects with missing data are excluded.
Investigated strategies

- Standard MI using logistic regression
  - Fitting logistic regression using observed data
  - Construct the posterior predictive distribution of the parameters
  - Fit new logistic regression using parameters simulated from the above posterior distribution to impute missing values
Investigated strategies

- Standard MI using Markov chain Monte Carlo method
  - Draw pseudo random samples from a target probability distribution
    \[
    \Pr(Y_{mis} \mid Y_{obs}) = \int \Pr(Y_{mis} \mid Y_{obs}, \theta) \Pr(\theta \mid Y_{obs}) d\theta
    \]
    where \( Y_{mis} \) represents the missing data
    \( Y_{obs} \) represents the observed data
    \( \theta \) represents the unknown parameters
Investigated strategies

- **Within-cluster MI**
  - Logistic regression
  - Markov chain Monte Carlo (MCMC) method

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Patient ID</th>
<th>Outcome</th>
<th>Blood Pressure</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>0</td>
<td>65</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>102</td>
<td>0</td>
<td>68</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>103</td>
<td>x</td>
<td>78</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>104</td>
<td>1</td>
<td>70</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>105</td>
<td>0</td>
<td>69</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>106</td>
<td>x</td>
<td>82</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>107</td>
<td>1</td>
<td>67</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>108</td>
<td>1</td>
<td>71</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>281</td>
<td>0</td>
<td>80</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>282</td>
<td>x</td>
<td>77</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>283</td>
<td>x</td>
<td>73</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>284</td>
<td>1</td>
<td>69</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>285</td>
<td>1</td>
<td>79</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>286</td>
<td>1</td>
<td>70</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>287</td>
<td>1</td>
<td>75</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>288</td>
<td>x</td>
<td>81</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>
Investigated strategies

- MI using cluster as a fixed effect
  - Fitting logistic regression using observed data
    - including cluster as a covariate
  - Construct the posterior predictive distribution of the parameters
  - Fit new logistic regression using parameters simulated from the above posterior distribution to impute missing values
Simulation Study

- Assumptions
  - Missing is at random.
  - Completely randomized CRTs with balanced design.
  - Two level of clustering.
  - Statistical analysis using generalized estimating equations (GEE) approach.
## Simulation Study

- Design of CRTs considered in simulation

<table>
<thead>
<tr>
<th>Type of CRT</th>
<th>Num. of clusters per arm</th>
<th>Num. of subjects per cluster</th>
<th>Intra-cluster correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small num. of cluster and large num. of subjects in each cluster</td>
<td>5</td>
<td>50, 100, 500</td>
<td>0.001, 0.01, 0.05</td>
</tr>
<tr>
<td>Large num. of cluster and small num. of subjects in each cluster</td>
<td>20</td>
<td>3, 10, 50</td>
<td>0.01, 0.1, 0.3</td>
</tr>
</tbody>
</table>
Simulation Study

- Generate outcomes for each cluster

\[ Y \sim \text{Binomial}(n, p) \]
\[ P \sim \text{Beta}(\alpha, \beta) \]
\[ \alpha = \pi \frac{1 - \rho}{\rho}; \quad \beta = (1 - \pi) \left( \frac{1 - \rho}{\rho} \right) \]

where \( Y \) is the outcome variable

- \( n \) is the number of subjects per cluster,
- \( \rho \) is ICC,
- \( \pi \) is the marginal probability of event,
- \( \pi = 0.4 \) for control group and 0.3 for intervention group.
Simulation Study

- Evaluation of the performance
  - Standardized bias
  - Root mean square error (RMSE)
  - Coverage probability
  - Average standard error of treatment effect
Results

Average Standard Error for CRT with 5 Clusters/Arm, 50 Subjects/Cluster, ICC=0.01, VIF=1.49

Percentage of Missing Data

Average Standard Error

- Complete Case Analysis
- Standard MI (Logistic)
- Standard MI (MCMC)
- Within Cluster MI (Logistic)
- Within Cluster MI (MCMC)
- MI (Logistic Fixed Effects)
Results

Average Standard Error for CRT with 5 Clusters/Arm, 100 Subjects/Cluster, ICC=0.05, VIF=5.95

Percentage of Missing Data:
- no missing
- missing 5%
- missing 15%
- missing 30%

Average Standard Error:
- Complete Case Analysis
- Standard MI (Logistic)
- Standard MI (MCMC)
- Within Cluster MI (Logistic)
- Within Cluster MI (MCMC)
- MI (Logistic Fixed Effects)
Results

Average Standard Error for CRT with 5 Clusters/Arm, 500 Subjects/Cluster, ICC=0.05, VIF=25.95
Results

Average Standard Error for CRT with 20 Clusters/Arm, 3 Subjects/Cluster, ICC=0.3, VIF=1.6

<table>
<thead>
<tr>
<th>Percentage of Missing Data</th>
<th>Complete Case Analysis</th>
<th>Standard MI (Logistic)</th>
<th>Standard MI (MCMC)</th>
<th>Within Cluster MI (Logistic)</th>
<th>Within Cluster MI (MCMC)</th>
<th>MI (Logistic Fixed Effects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing 15%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing 30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Average Standard Error for CRT with 20 Clusters/Arm, 10 Subjects/Cluster, ICC=0.3, VIF=6.7

Percentage of Missing Data:
- no missing
- missing 5%
- missing 15%
- missing 30%

Average Standard Error:
- Complete Case Analysis
- Standard MI (Logistic)
- Standard MI (MCMC)
- Within Cluster MI (Logistic)
- Within Cluster MI (MCMC)
- MI (Logistic Fixed Effects)
Results

Average Standard Error for CRT with 20 Clusters/Arm, 50 Subject/Cluster, ICC=0.3, VIF=15.7
Conclusions

- Determinants for selecting an appropriate missing data strategy
  - Percentage of missing data
  - Num. of clusters per arm
  - Num. of subjects per cluster
  - ICC
  - VIF
Conclusions

- Complete case analysis can be used when
  - Percentage of missing data is small (<10%)
  - Design effect is large (VIF>6)
Conclusions

- Standard MI using logistic regression or MCMC method
  - tends to underestimate the standard error of the treatment effect
  - can be used to impute the missing values when the percentage of missing data is small (<15%) and the design effect is small (VIF≤3)
  - leads to severe underestimate of the standard error of the treatment effect when percentage of missing data and design effect is large
Conclusions

- Within-cluster MI using logistic regression or MCMC method
  - may not work for CRTs with SL design
  - tends to overestimate the standard error of the treatment effect
  - can be used to impute missing data from CRTs with SL design, especially when design effect is large (VIF>3)
Conclusions

- MI using logistic regression with cluster as a fixed effect
  
  - substantially overestimates the standard error of the treatment effect for RCTs with SL design
  
  - overestimates standard error of the treatment effect for RCTs with LS design with extremely small number of subjects within each cluster
Limitations

- CRT design settings investigated
  - Completely randomized design
  - Two level of clustering
  - Balanced design:
    - equal number of clusters per arm
    - equal number of subjects per cluster

- Other imputation strategies may be valid but not investigated: propensity score method
Thanks for your attention!