# 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

| Study | Taljaard (2008) | Andrige (2011) |
|---|---|---|
| Assumption | Missing completely at random | Missing completely at random and missing at random |
| Approach | Simulation | Simulation |
| Strategies Investigated | 1. Standard regression multiple imputation (MI) ignoring clusters<br>2. Cluster mean imputation<br>3. Within-cluster MI using approximate Bayesian Bootstrap (ABB) method<br>4. Pooled MI using ABB method<br>5. Mixed–effects regression MI | 1. Fixed-effects MI incorporating fixed effects for cluster |

1. Taljaard M, Donner A, Klar N. (2008) Imputation strategies for missing continuous outcomes in cluster randomized trials. Biom J 50(3): 329-345.
2. Andridge RR. (2011) Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. Biom J 53(1): 57-74.

# Missing Data Strategies

- Missing binary outcomes in CRTs

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

Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, CHAT investigators. (2011) Imputation strategies for missing binary outcomes in cluster randomized trials. BMC Med Res Methodol 11: 18.

# 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

$$\Pr(Y_{mis}|Y_{obs}) = \int \Pr(Y_{mis} \mid Y_{obs}, \theta) \Pr(\theta \mid Y_{obs}) d\theta$$

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

| Cluster | Patient ID | Outcome | Potential Predictors | |
|---|---|---|---|---|
| | | Blood Pressure | Age | Sex |
| 1 | 101 | 0 | 65 | F |
| 1 | 102 | 0 | 68 | F |
| 1 | 103 | x | 78 | M |
| 1 | 104 | 1 | 70 | M |
| 1 | 105 | 0 | 69 | M |
| 1 | 106 | x | 82 | F |
| 1 | 107 | 1 | 67 | M |
| 1 | 108 | 1 | 71 | M |
| | ..... | | | |
| 28 | 281 | 0 | 80 | M |
| 28 | 282 | x | 77 | F |
| 28 | 283 | x | 73 | F |
| 28 | 284 | 1 | 69 | F |
| 28 | 285 | 1 | 79 | M |
| 28 | 286 | 1 | 70 | F |
| 28 | 287 | 1 | 75 | F |
| 28 | 288 | x | 81 | F |

# 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

| Type of CRT | Num. of clusters per arm | Num. of subjects per cluster | Intra-cluster correlation coefficient |
|---|---|---|---|
| **Small num. of cluster and large num. of subjects in each cluster** | **5** | **50**<br>**100**<br>**500** | **0.001**<br>**0.01**<br>**0.05** |
| **Large num. of cluster and small num. of subjects in each cluster** | **20** | **3**<br>**10**<br>**50** | **0.01**<br>**0.1**<br>**0.3** |

# Simulation Study

- Generate outcomes for each cluster

$$Y \sim Binomial(n, p)$$

$$P \sim Beta(\alpha, \beta)$$

$$\alpha = \pi \frac{1-\rho}{\rho}; \quad \beta = (1-\pi)\left(\frac{1-\rho}{\rho}\right)$$

where $Y$ is the outcome variable

$n$ is the number of subjects per cluster,

$\rho$ is ICC,

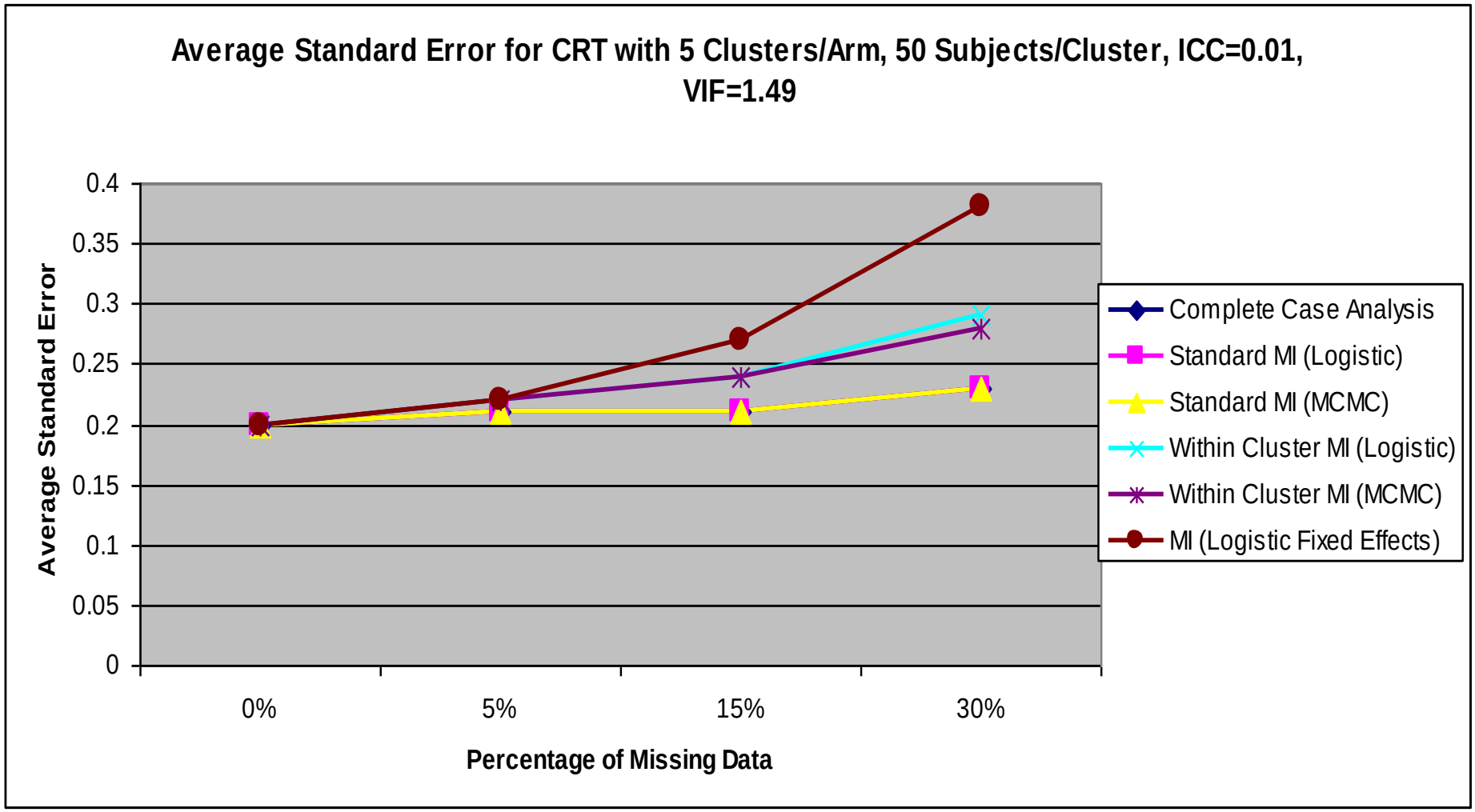
$\pi$ is the marginal probability of event,

$\pi = 0.4$ for control group and $0.3$ for intervention group.

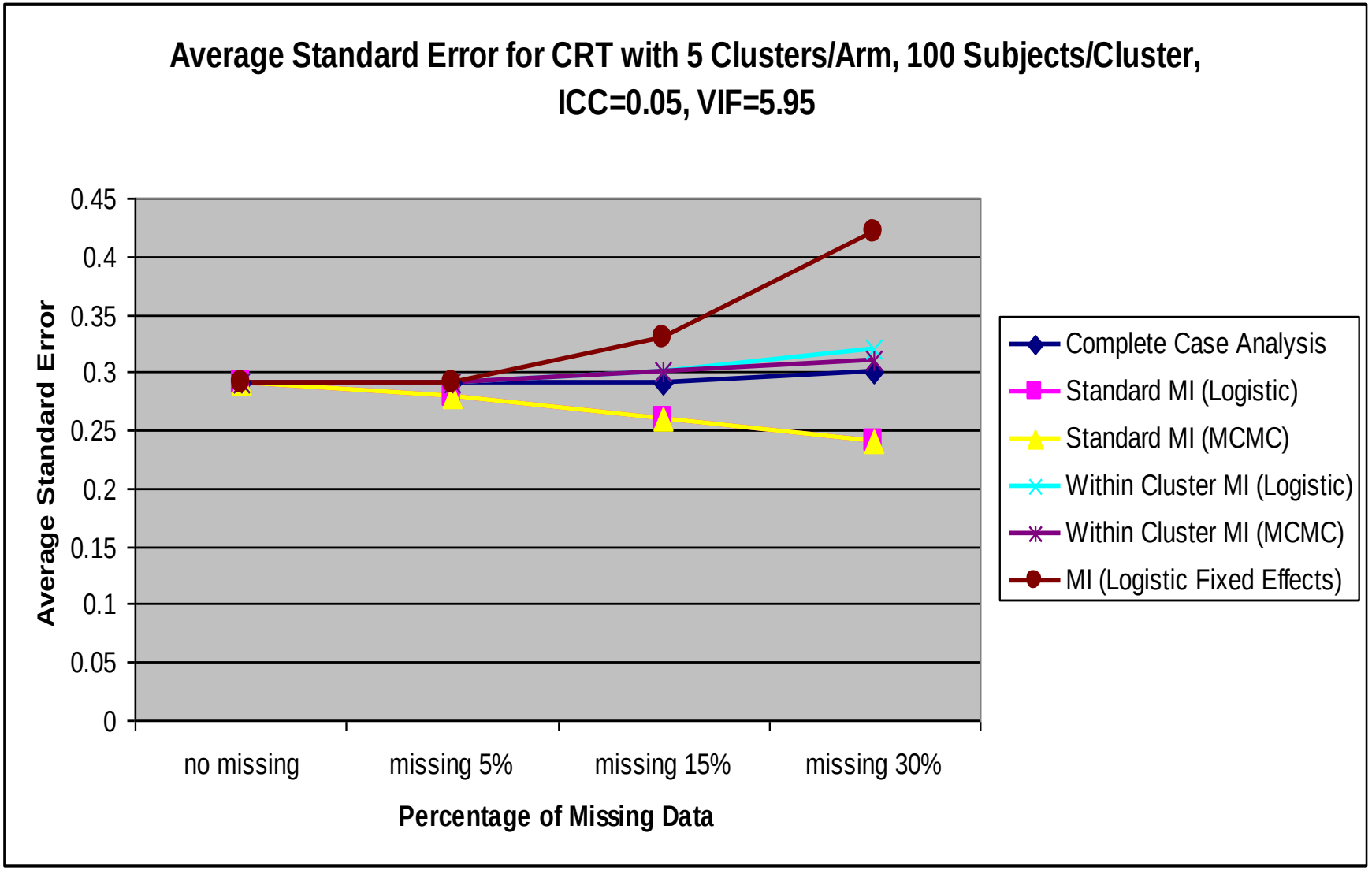
# Simulation Study

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

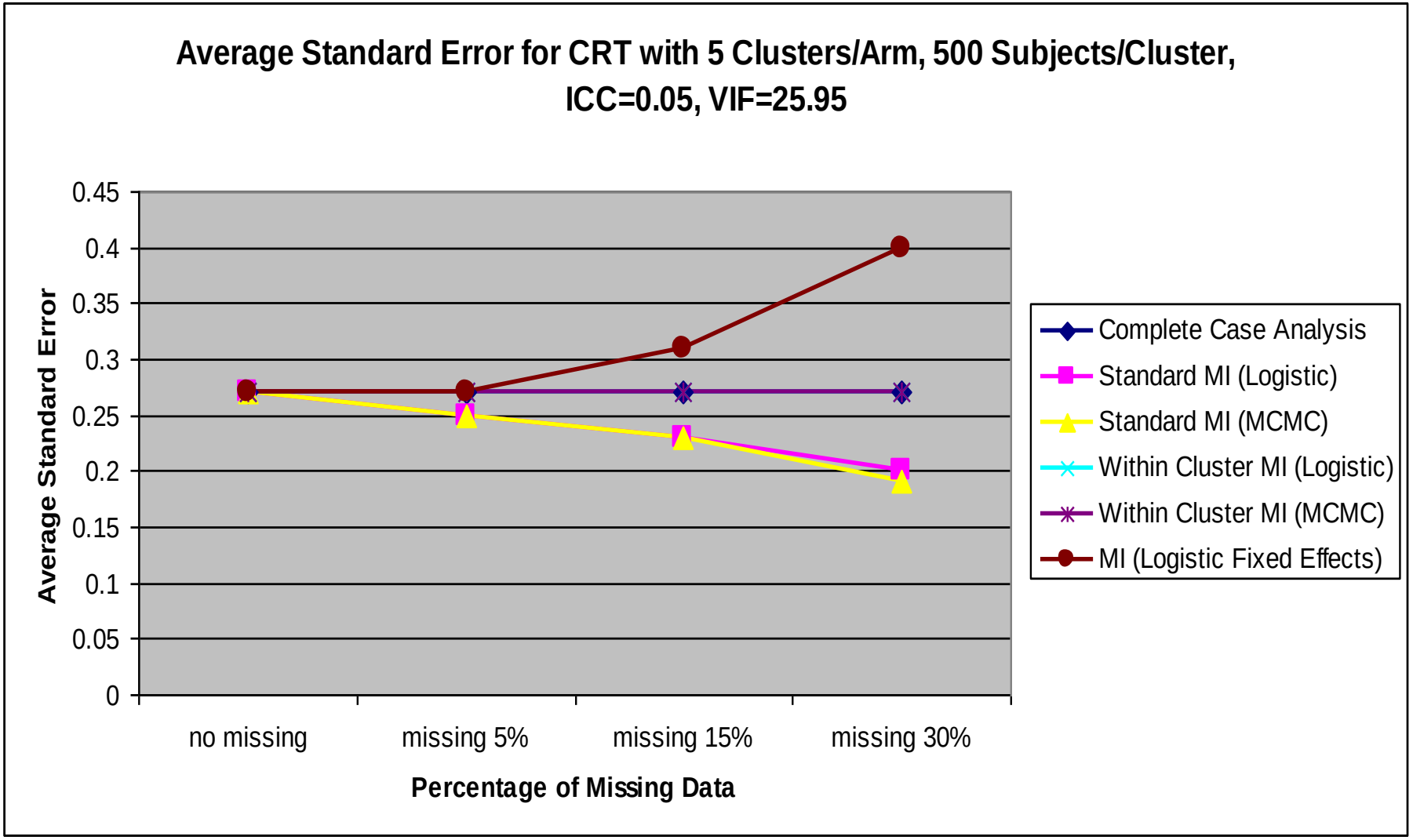
# Results


Average Standard Error for CRT with 5 Clusters/Arm, 50 Subjects/Cluster, ICC=0.01, VIF=1.49
Average Standard Error
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0%
5%
15%
30%
Percentage of Missing Data
Complete Case Analysis
Standard MI (Logistic)
Standard MI (MCMC)
Within Cluster MI (Logistic)
Within Cluster MI (MCMC)
MI (Logistic Fixed Effects)

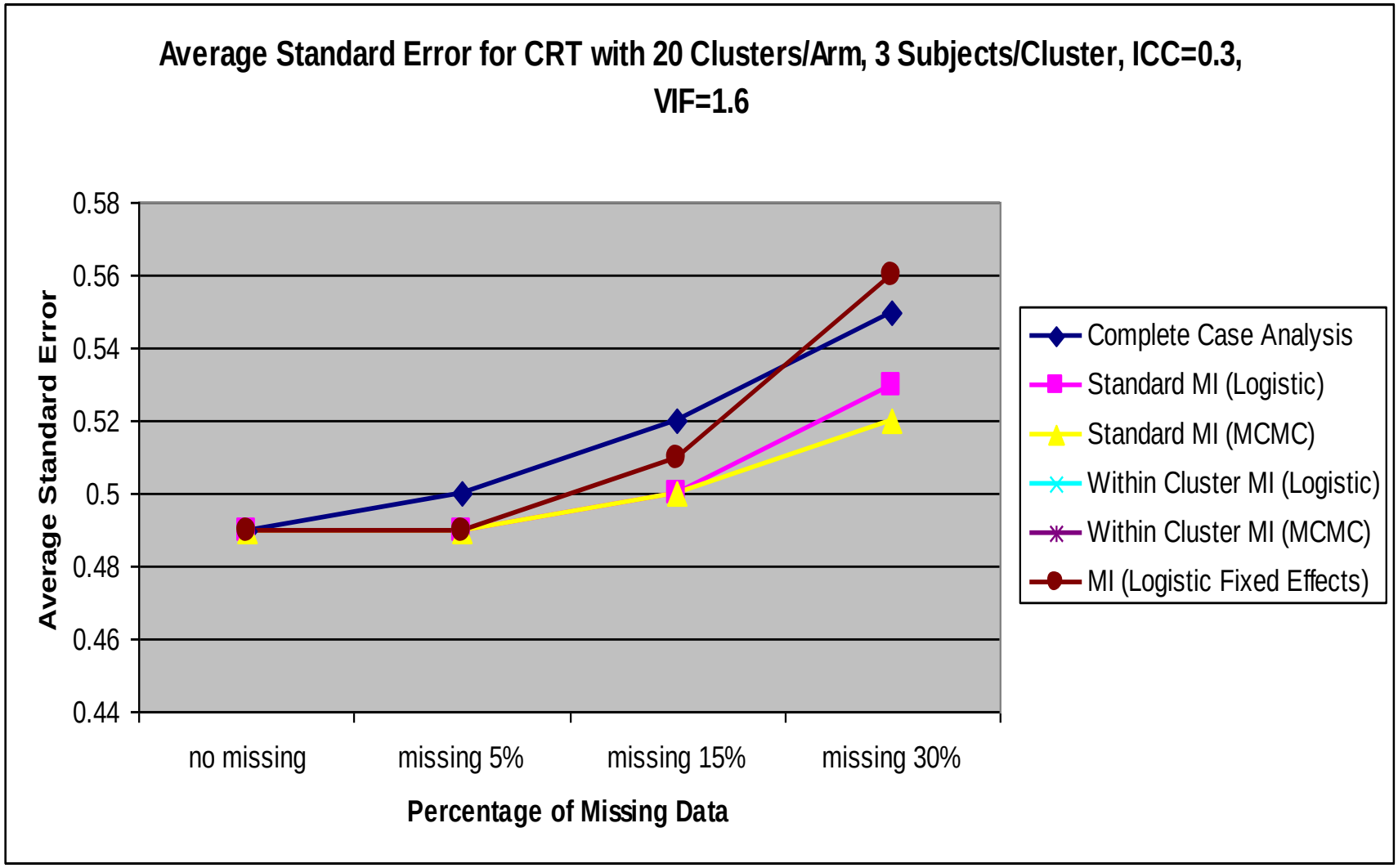
# Results


Average Standard Error for CRT with 5 Clusters/Arm, 100 Subjects/Cluster, ICC=0.05, VIF=5.95
Average Standard Error
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
no missing
missing 5%
missing 15%
missing 30%
Percentage of Missing Data
Complete Case Analysis
Standard MI (Logistic)
Standard MI (MCMC)
Within Cluster MI (Logistic)
Within Cluster MI (MCMC)
MI (Logistic Fixed Effects)

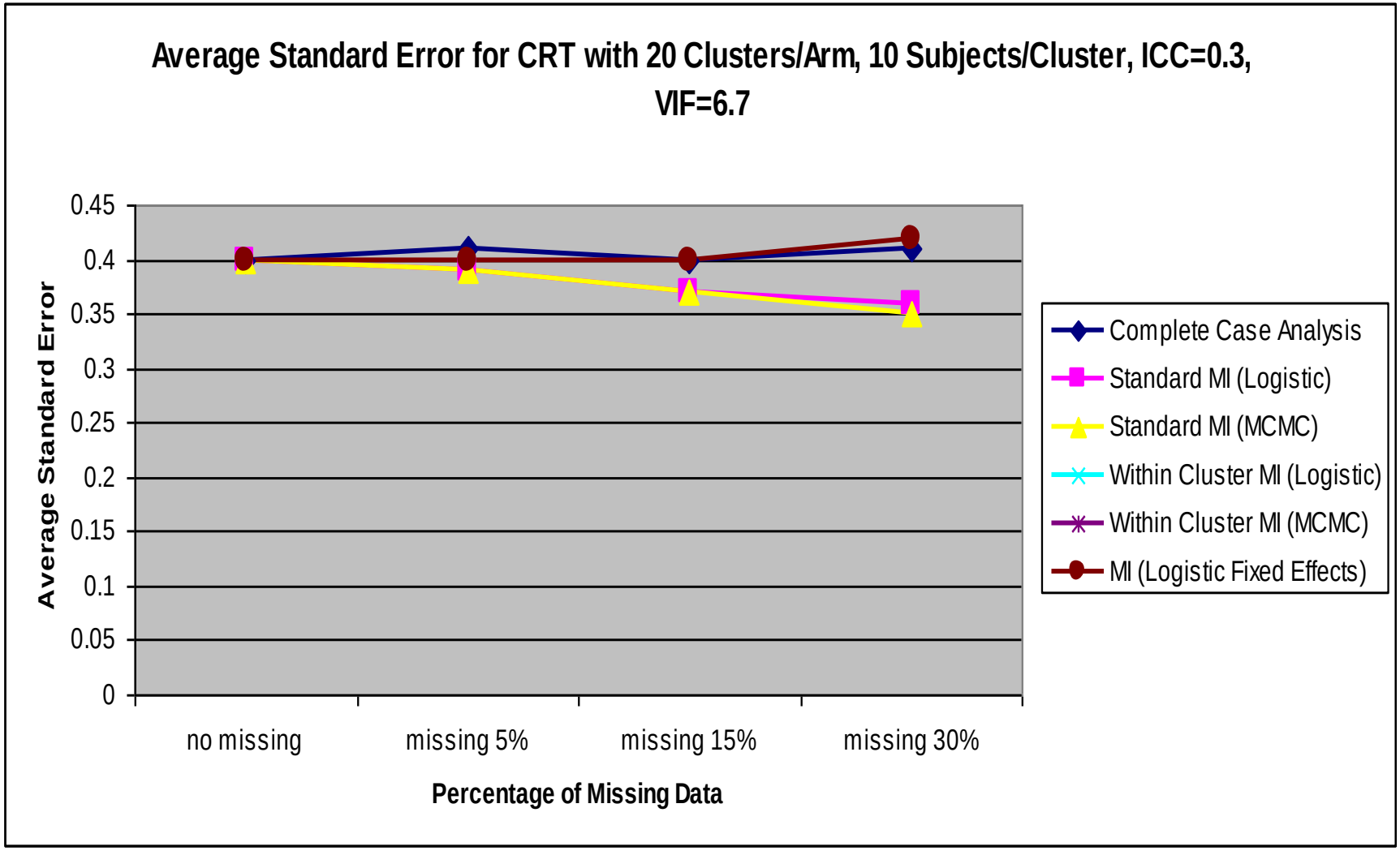
# Results


Average Standard Error for CRT with 5 Clusters/Arm, 500 Subjects/Cluster, ICC=0.05, VIF=25.95
Average Standard Error
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
no missing
missing 5%
missing 15%
missing 30%
Percentage of Missing Data
Complete Case Analysis
Standard MI (Logistic)
Standard MI (MCMC)
Within Cluster MI (Logistic)
Within Cluster MI (MCMC)
MI (Logistic Fixed Effects)

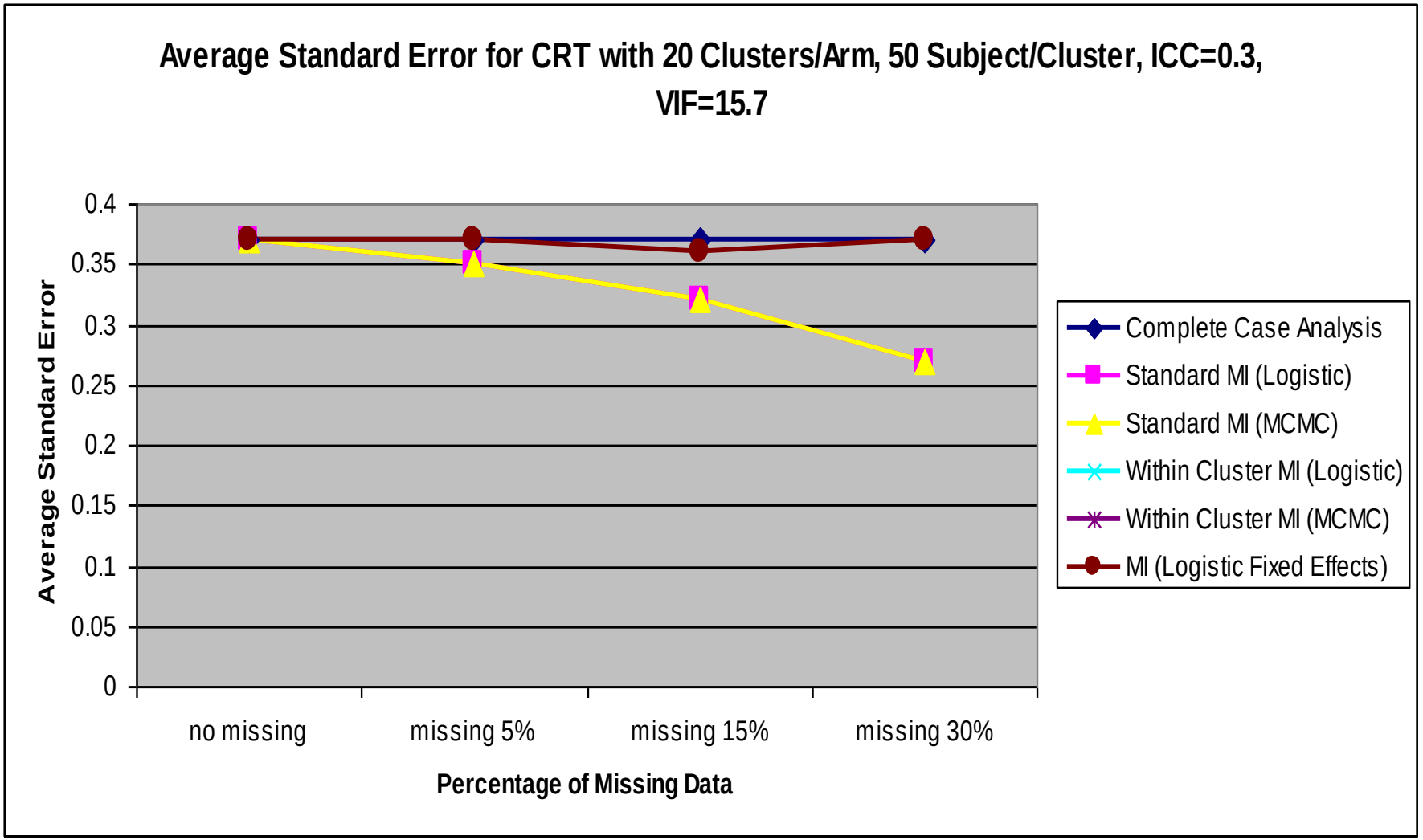
# Results


Average Standard Error for CRT with 20 Clusters/Arm, 3 Subjects/Cluster, ICC=0.3, VIF=1.6
Average Standard Error
0.58
0.56
0.54
0.52
0.5
0.48
0.46
0.44
no missing
missing 5%
missing 15%
missing 30%
Percentage of Missing Data
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, 10 Subjects/Cluster, ICC=0.3, VIF=6.7
Average Standard Error
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
no missing
missing 5%
missing 15%
missing 30%
Percentage of Missing Data
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
Average Standard Error
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
no missing
missing 5%
missing 15%
missing 30%
Percentage of Missing Data
Complete Case Analysis
Standard MI (Logistic)
Standard MI (MCMC)
Within Cluster MI (Logistic)
Within Cluster MI (MCMC)
MI (Logistic Fixed Effects)

# 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!