A Comparison of Missing Data Methods for Clinical Trials

SUDS Scholar: Bilin Nong

Supervisors: Prof. Kuan Liu, Prof. Kevin Thorpe

Dalla Lana School of Public Health, University of Toronto

UNIVERSITY OF TORONTO

DATA SCIENCES INSTITUTE

Background

- Missing data emerges in various ways in randomized controlled trials (RCTs), such as patient withdrawal.
- Improper handling of missing data may lead to bias and reduced precision in estimating treatment effect.
- There has not been a comprehensive evaluation of the relative performance of missing data methods for RCTs.

Objectives

- Evaluate the performance of common missing data handling approaches under different missing data mechanisms via statistical simulations.
- Provide practical suggestions and recommendations on dealing with missingness in RCTs.

Methods

Simulation Settings

- We used R statistical software to simulate a trial with continuous outcomes under a linear regression model:

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 z_i + \epsilon_i$$

  - The random error: $\epsilon_i \sim N(0,10)$
  - The intercept: $\beta_0 = 0$
  - The coefficients: $\beta_1 = \beta_2 = \beta_3 = 1$
  - All x variables correlates with each other with a correlation = 0.5
  - The treatment effect: β_4 is independent of all x variables
- Sample size: n = 200, 500
- We ran 1000 simulations for each trial

Missing Data Methods & Analysis

We applied common missing data approaches to each simulated trial, including:

- **complete-case analysis (CC)**
- **Regression Imputation (RI):** We imputed missing values with the predicted values from a regression model.
- **Multiple Imputations (MI):** We imputed multiple completed datasets using multiple imputation by chained equations and then pools the treatment effect estimates from each imputed dataset using the Rubin's rule.
- **Inverse Probability Weighting (IPW):** We used logistic regression and BART to estimate the probability of being observed and then estimate the treatment effect using a weighted linear regression.
- We studied both **unadjusted** (no adjustment of x) & **adjusted treatment effect** under each simulated trial.

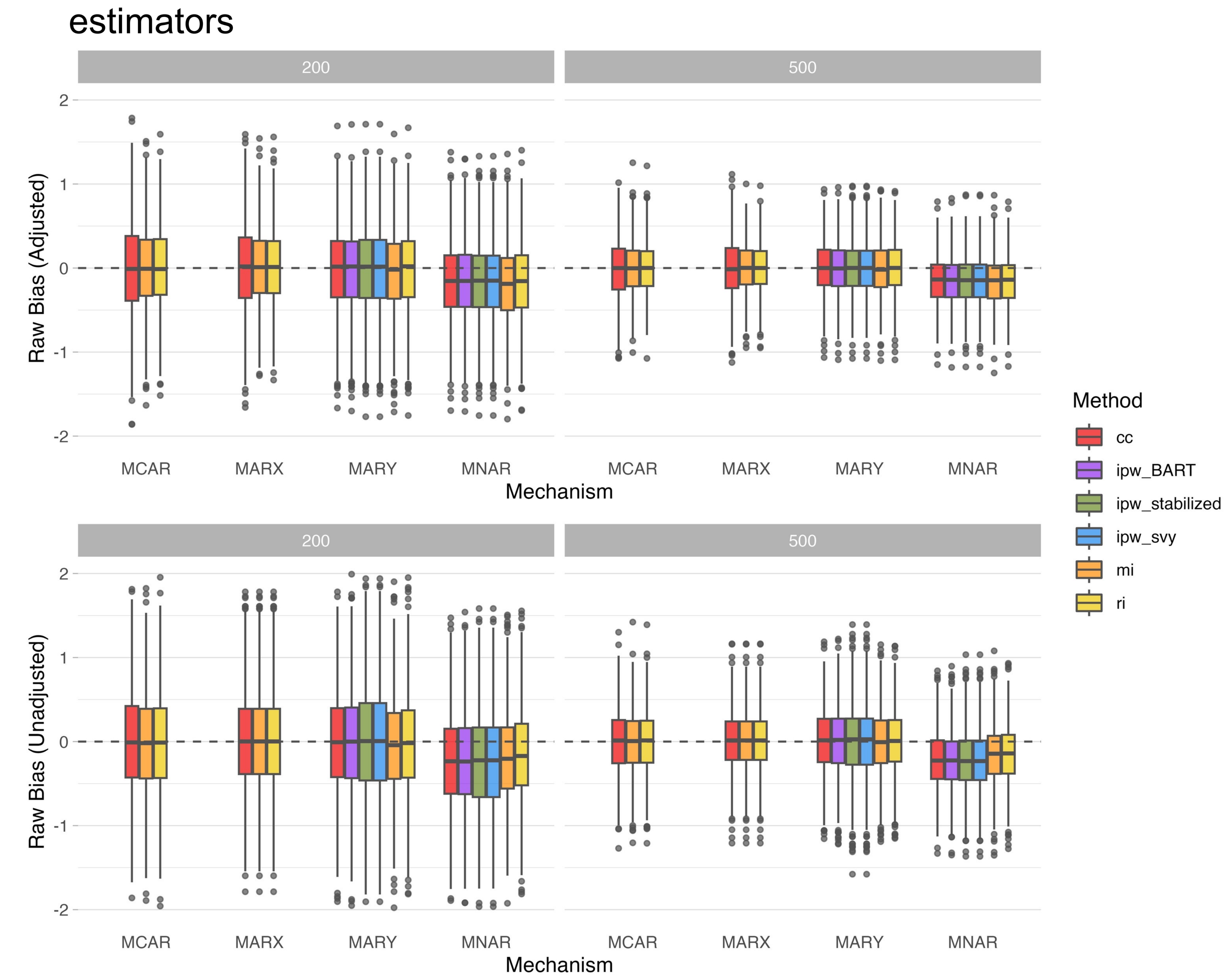
Missing Mechanisms

We simulated trials contain missing data under various missing mechanisms:

- **Missing Completely at Random (MCAR)**: 10% prob. of missing for x variables and 15% for y variable.
- **Missing at Random (MAR)**
  - **MARX:** x variables contains missingness, prob. of missing is predicted by other covariates.
  - **MARY:** y variable contains missingness, prob. of missing is predicted by all 3 covariates.
- **Missing Not at Random (MNAR):** the higher probability of missing for y values that are further away from it's mean.

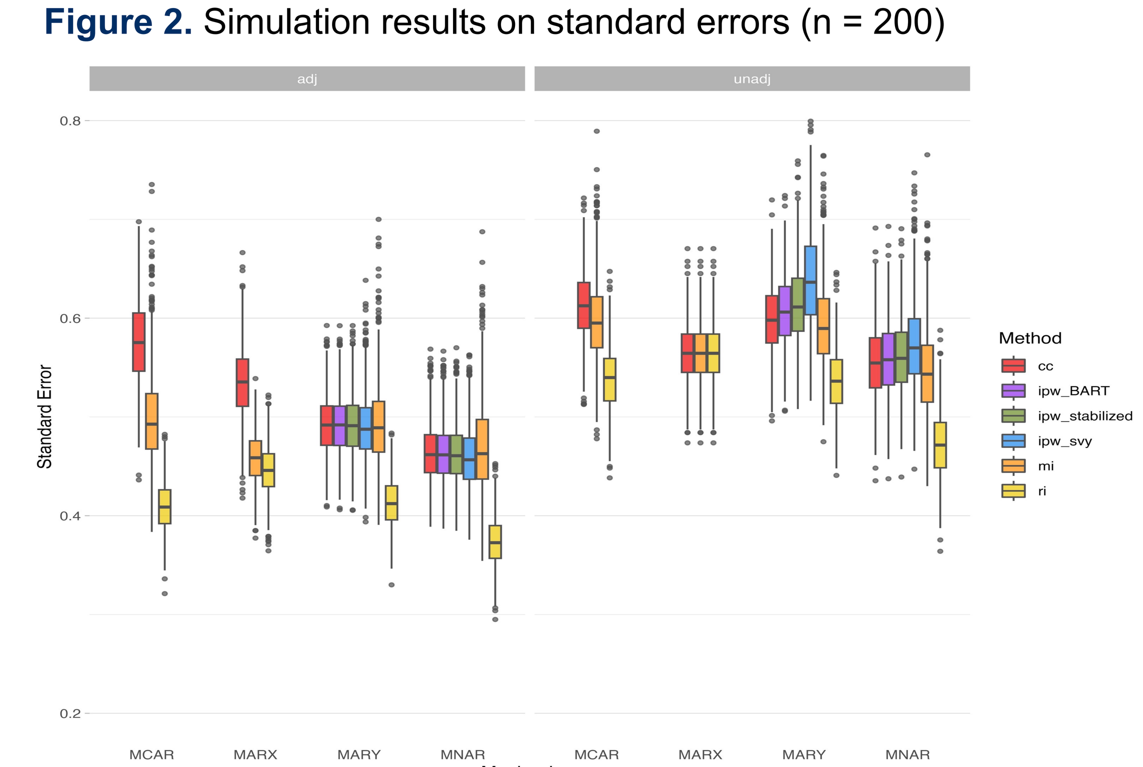
Results

Figure 1. Simulation results on the bias distribution of treatment effect estimators



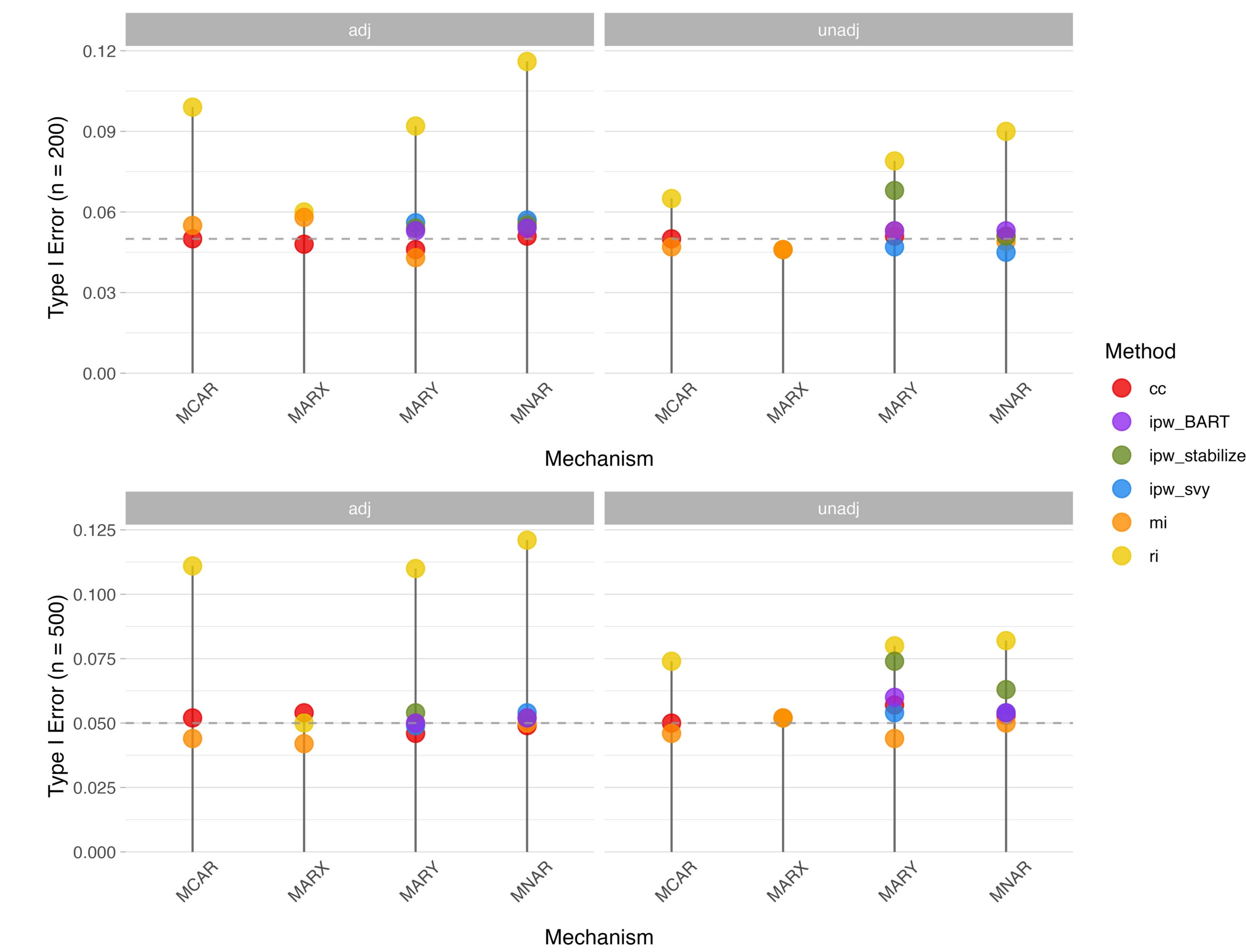
- The bias of all methods are comparable.
- The adjusted analysis yielded less biased result than the unadjusted analysis.
- MNAR: no methods used in this study corrected the bias.

Figure 2. Simulation results on standard errors (n = 200)



- Unadjusted analysis yielded higher standard error than adjusted analysis.
- RI returned the lowest standard errors under all missing mechanisms.
- MI yielded the largest standard errors under MARY and MNAR than other methods.

Figure 3. Type I Error lollipop plots



- RI method returned the highest type I Error.
- Other approaches achieved close to nominal type I error rate of 0.05.

Conclusion & Outlook

- under MCAR: Imputation methods are not needed to correct bias.
- under MARX: Imputation methods and complete case analysis yielded comparable bias.
- under MARY: Multiple imputations is not preferred due to large standard error. In contrast, weighting methods work well.
- under MNAR: none of the methods worked well without knowing the correct missing model!
- Address the complexity of the MNAR scenario!
- RCTs with binary outcomes & survival outcomes!

References

1. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol. 2018;18:168.
2. Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research, 22(3), 278–295.
3. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statist Med. 2006;25:4279–92.
4. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer US; 2021.
5. Pocock SJ. Clinical Trials: A Practical Approach.
6. Molenberghs G. Missing Data in Clinical Studies.

R Packages

1. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). Welcome to the **tidyverse**. Journal of Open Source Software, 4(43), 1686.
2. Goldfeld K, Wujciak-Jens J (2020). "**simstudy**: Illuminating research methods through data generation." Journal of Open Source Software, 5(54), 2763.
3. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3), 1-67.
4. Microsoft, & Weston, S. (2019). **doParallel**: Foreach parallel adaptor for the 'parallel' package. R package version 1.0. 15.
5. Lumley T (2004). "Analysis of Complex **Survey** Samples." Journal of Statistical Software, 9(1), 1-19. R package verson 2.2.
6. Greenman, M., & Halpern-Manners, A. (2018). **WeightIt**: A software package for propensity score weighting. Journal of statistical software, 84(16), 1-25.