HANDLING MISSING DATA IN RANDOMIZATION-BASED INFERENCE

by

Xiao Tan

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

Dr. William F. Rosenberger, Dissertation Director

Dr. Clifton Sutton, Committee Member

Dr. Pramita Bagchi, Committee Member

Dr. Diane Uschner, Committee Member

Dr. Jiayang Sun, Department Chair

Date: 04 April 2022

Spring Semester 2022
George Mason University
Fairfax, VA

Handling Missing Data in Randomization-Based Inference

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Xiao Tan
Master of Arts
University of California, Santa Barbara, 2017
Bachelor of Science
Harbin University of Commerce, 2015

Director: Dr. William F. Rosenberger, Professor
Department of Statistics

Spring Semester 2022
George Mason University
Fairfax, VA

Dedication

To my parents, Zhizhong Tan and Jian Gong, who always support me in my whole life, my fiance, Hongfei Du, who always comforts me when I was experiencing mental or intellectual struggles, and my dear friends, who brings me joy and encouragement. May health and happiness always surround them.

Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. William F. Rosenberger, for his invaluable guidance, continued support, and endless patience throughout my Ph.D. life. He is an extraordinary supervisor with an insightful vision, and I feel that I am incredibly fortunate to have the opportunity to finish my Ph.D. under his mentorship. This dissertation would not have been possible without his guidance during my graduate education.

Besides, I extend my gratitude to other committee members for their constant support throughout my research. I sincerely appreciate Dr. Diane Uschner for her innovative ideas and fantastic suggestions on my dissertation. I would like to thank Dr. Clifton Sutton for teaching me how to apply statistical methods in practice when working in the consulting center. My thanks go to Dr. Pramita Bagchi, a great committee member and instructor who is patient with every question I have during her office hours.

I also owe a debt to the faculty and staff members in the Statistics Department. I would like to express my gratitude to Dr. Yanying Wang, a great statistician and friend, who kindly helped me with my questions. Specifically, I would like to thank Dr. Anand Y. Vidyashankar, Dr. Scott Bruce, Dr. Guoqing Diao, Dr. Daniel B. Carr, Dr. Wanli Qiao, and Dr. Keith Crank, who kindly taught me statistics knowledge and shared insightful ideas during my four years at GMU. Also, I would like to thank Dr. Martin Slawski, a great Ph.D. program director, who always responds to my questions with patience and gives me insightful feedbacks on my presentation at student seminars. Further, I am grateful to Dr. Brett D. Hunter, Mrs. Mitchell, and Carroll for helping me with my questions, as well as their kindness and patience.

Last, I would like to express my gratitude to the GMU for their generous funding support for my whole Ph.D. life. I would not finish this Ph.D. degree without their generous support. I sincerely appreciate that GMU offers me the opportunity to receive education in such an exciting field, and the platform to communicate and collaborate with these fantastic people in the Statistics Department.

Table of Contents

List of Tables

List of Figures

Abstract

HANDLING MISSING DATA IN RANDOMIZATION-BASED INFERENCE

Xiao Tan, PhD

George Mason University, 2022

Dissertation Director: Dr. William F. Rosenberger

Randomized controlled trials (RCTs) serve as the gold standard in researching and developing new therapeutics. A new treatment's effectiveness is evaluated by comparing it to existing or standard treatment in an RCT. However, the imbalance in participants' characteristics between groups would harm such comparison. The act of randomization on patients mitigates the bias caused by such imbalance in the evaluation of treatment effects. Randomization-based inference was first introduced by Sir R.A. Fisher as an approach to evaluate treatment effects in an RCT. The limit in computing power has slowed its development in the past. However, the tremendous growth of computing technology enables us to compute randomization tests easily.

Randomization-based inference is a natural way to analyze data from a clinical trial. But the presence of missing outcome data is problematic: if the data are removed, the randomization distribution is destroyed, and randomization tests have no validity. There are no randomization-based methods to handle missing data. In this thesis, the unconditional reference set method, the conditional reference set method and the randomization-based multiple imputation are described to handle missingness while preserving the randomization distribution.

Randomization-based missing data methods are compared to population-based and parametric imputation approaches via the metrics of type I error rates and power under both homogeneous and heterogeneous population models. Randomization-based analogs of standard missing data mechanisms are described, and a randomization-based procedure is proposed to determine if data are missing completely at random. A large simulation protocol is implemented to conclude that the unconditional, the conditional reference sets method and the randomization-based multiple imputation are reasonable approaches to handle missing data in patients' missingness in the context of a two-armed RCT.

Chapter 1: Background and Significance

1.1 Introduction

The objective of most clinical trials is providing an unbiased comparison between treatments. However, missing data can seriously compromise the comparison between treatments from a statistical inference perspective, introducing bias in the comparison and diminishing the statistical power of a study. Usually patients enter a study sequentially and are randomized to one of two or more treatments. It is not unusual to see patients drop out before the end of a study. This may because patients who are assigned to the experimental treatment may choose to discontinue the medication when they feel better, and patients on placebo may also stop taking medication because they think the treatment is non-effective during the study, or patients may be lost to follow-up for reasons unrelated to the trial. Different missing data mechanisms affect the choice of statistical inference methods when dealing with data with missingness. Unfortunately, determining the underlying missing data mechanism is hard in practice, making data analysis with missingness even more complicated.

Even when there is no missingness in the data from randomized clinical trials, there is a current debate over the choice of analysis methods: population-based methods and randomization-based methods. For instance, if we are interested in comparing the treatment effects of two-armed randomized clinical trials with experimental and placebo treatments, assuming patients' responses are continuous, it is natural to use a population-based method such as a two-sample t-test. However, the random sampling assumption for the t-test itself is questionable. As Rosenberger et al. (2019) point out, clinical trials are designed experiments, and there are no existing populations of patients taking experimental medications or procedures. The clinical trial creates a finite population consisting of patients

who are involved in the trial. Hence there is no random sampling. The initial assumption for population-based methods, i.e., random sampling, is inappropriate. However, the randomization-based inference is a more natural and objective method to investigate different treatments' efficacy. Statistical inference using a randomization test is based on the probability distribution derived from the randomization procedures, which are adopted in the trials without any random sampling assumptions on patients' collection. Randomization tests preserve the type I error rates, even under heterogeneity. More details about the randomization test are discussed in a later section.

The randomization test's virtues are explained by Rosenberger et al. (2019) and these motivate us to implement it in practice. Nevertheless, the lack of reliable randomization-based methods adjusted for missing data can hinder the use of the randomization test, especially when there are many population-based methods available that handle data with missingness, such as the direct likelihood and multiple imputation. Confronting this situation, a randomization-based method adjusting for the missing values is desired. We focus on the situations when the missingness only occurs in patients' responses, not treatment assignments, after they participate in a two-armed randomized clinical trial. The methods we discuss are versatile when dealing with different types of responses. The continuous and binary cases are both well-investigated in the simulation studies. We evaluate the methods' performance under different missing data mechanisms. The new method's validity is also tested under situations where there is heterogeneity in the patient's outcomes.

1.2 Randomization Procedures

Before discussing randomization procedures in detail, it is first necessary to clarify our goal in a randomized clinical trial. Assume a new treatment is available, and we wish to compare its efficacy with an existing treatment, or placebo. Patients are recruited into the trial, and two treatments are assigned to patients according to a randomization procedure. Procedures promoting the comparability between two groups of patients who receive different treatments and allocating treatments to patients without introducing human factors are

essential to protect the comparison from bias. Randomization provides these protections. We focus on reviewing the classical randomization procedures in the following.

Let $\boldsymbol{T} = T_1, \ldots, T_n$ be a vector of random treatment assignments for n patients, where $T_j = 1$ if the patient j received treatment A and $T_j = 0$ if received treatment B.

1.2.1 Complete Randomization

For complete randomization (CR), the treatment assignments $T_1, \ldots, T_n$ are independent and identically distributed Bernoulli random variables where $P(T_j = 1) = 1/2$ for $j = 1, \ldots, n$. CR is easy to implement in practice. However, this procedure is less attractive since there is a non-negligible probability of obtaining a severe imbalance in treatment assignments, i.e., significantly more patients receive one treatment compared to another treatment. Note that all possible treatment assignment sequences are equiprobable. Rosenberger and Lachin (2016) discuss this imbalance with simulation studies and theoretical results. Some forced-balance procedures are proposed to deal with the imbalance. Forced-balance procedures result in exactly $n/2$ patients allocate to treatment A and $n/2$ patients to treatment B.

1.2.2 Random Allocation rule

Unlike CR, the random allocation rule (RAR) relies on the previous treatment assignments of $j-1$ patients when allocating a treatment to the jth patient. Let $\mathcal{F}_n$ be the treatment assignments for first n patients, where $\mathcal{F}_n = \{T_1, \ldots, T_n\}$. The RAR is defined by the following rule - the probability that the jth patient receives treatment A is given by

$$E(T_j|\mathcal{F}_{j-1}) = \frac{\frac{n}{2} - N_A(j-1)}{n-(j-1)}, \quad j = 2, \ldots, n$$

where $P(T_1) = 1/2$ and $N_A(j-1)$ is the number of patients who received treatment A after $j-1$ patients have been assigned. The treatment assignments are predictable at some stages in the trial, which may result in selection bias. For instance, if $n/2$ patients have

already received treatment A, the remaining patients who have not received treatments must receive treatment B; thus, a significant imbalance also occurs midway through the trial. The imbalance makes the comparison between treatments questionable, especially when there is a time trend in patients' responses. The influences of imbalance under the RAR are discussed for large n by Rosenberger and Lachin (2016).

1.2.3 Truncated Binomial Design

Blackwell and Hodges (1957) proposed the truncated binomial design (TBD). The TBD allocates exactly $n/2$ patients to each treatment when implementing a two-armed randomized clinical trial. The rule for allocating patients under the TBD is summarized as the following, when n is even:

$$\begin{aligned} E(T_j|\mathcal{F}_{j-1}) &= \frac{1}{2}, && \text{if } \max\{N_A(j-1), N_B(j-1)\} < n/2 \\ &= 0, && \text{if } N_A(j-1) = n/2, \\ &= 1, && \text{if } N_A(j-1) = n/2. \end{aligned}$$

Similar to the problems in the RAR, the randomization sequences are predictable at some stages, and imbalances in the assignments are expected to occur during the trial.

1.2.4 Permuted Block Design

Severe imbalances may occur in the course of the trial if the designs discussed above are adopted. However, the permuted block design (PBD) better controls possible adverse effects due to imbalances, compared to CR, the RAR, and the TBD. We see how this works by clarifying the rule for the PBD. For PBD, M_B blocks are established, and each block has $m_B = n/M_B$ patients. Assume both M_B and m_B are even positive integers. Within each block, $m_B/2$ patients receive one treatment, and the rest receive another treatment. The RAR or the TBD are implemented when allocating the patients within each block.

Following this idea, it is easy to see that the maximum value for the possible imbalance is $m_B/2$ during the trial. In practice, block sizes should be greater than two. By introducing blocks, the extent of the imbalance alleviates. However, selection bias can not be prevented because of the predictability in the randomization sequences.

1.2.5 Random Block Design

Unlike the PBD, the block size varies in the random block design (RBD). The variability in block sizes reduces the adverse effects of selection bias on the randomization. Define B_j, $j = 1, \ldots, n$, as one half of the block size of the block containing the jth patient, so that B_j is a random variable from a discrete uniform distribution, and $B_{\max}$ is the largest possible value for B_j. The rule for the RBD is clearly defined in Rosenberger and Lachin (2016). The position of a jth patient within its block is defined as R_j, and this depends on the block size B_j. The RAR or the TBD were used to allocate treatments within a block. For instance, if the RAR was adopted, we have

$$E(T_j|\mathcal{F}_j, B_j, R_j) = \frac{\frac{B_j}{2} - \sum_{l=j+1-R_j}^{j-1} T_l}{B_j - R_j + 1}$$

There is a chance that the last block is unfilled since n usually is not known in advance.

1.2.6 Biased Coin Design

The biased coin design (BCD) was proposed by Efron (1971). The allocation rule is described as the following:

$$\begin{aligned} E(T_j|\mathcal{F}_{j-1}) &= \frac{1}{2}, && \text{if } D_{j-1} = 0, \\ &= \gamma, && \text{if } D_{j-1} < 0, \\ &= 1 - \gamma, && \text{if } D_{j-1} > 0, \end{aligned}$$

where D_n measures the differences of treatment assignments between treatment A and B (e.g., $D_n = N_A(n) - N_B(n) = 2N_A(n) - n$), and *gamma* is a constant, $\gamma \in (0.5, 1]$. Efron's original paper provides the recommendation of γ. In his paper, he states:

> The value $\gamma = 2/3$, which is the author's personal favourite, will be seen to yield generally good designs and will be featured in the numerical categories.

1.2.7 Big Stick Design

Soares and Wu (1983) introduced the big stick design (BSD). Define an imbalance tolerance parameter b, which is a positive integer. The level of imbalance is controlled within an acceptable range by b, which is fixed in advance. The rule for the BSD described as the following:

$$\begin{aligned} E(T_j|\mathcal{F}_{j-1}) &= \frac{1}{2}, && \text{if } |D_{j-1}| < b \\ &= 0, && \text{if } D_{j-1} = b \\ &= 1, && \text{if } D_{j-1} = -b. \end{aligned}$$

1.3 Randomization Test

Under a population model, to investigate the difference between treatments A and B, we compare the outcomes from two groups of patients using population-based methods, such as the t-test. For example, we assume patients in treatment A are a random sample from the population of patients taking treatment A. As we discussed in the introduction, the random sampling assumption is inappropriately adopted in a randomized clinical trial. The only randomness in the trial is induced by the implemented randomization procedure, which is ignored in the traditional analysis. Without a random sampling basis, population-based methods are inapplicable to data from the randomized clinical trial. However, the randomization test is the remedy for this. In Rosenberger and Lachin (2016), this philosophy

was summarized as follows:

> Fortunately, the use of randomization provides the basis for an assumption-free statistical test of the equality of the treatments among the n patients actually enrolled and studied. These are known as randomization tests.

For population-based methods, in the null hypothesis, we usually assume equality in the parameters from populations taking two different treatments A and B. In randomization tests, for the null hypothesis, we assume that the assignments of treatment A and B do not affect patients' responses. Rosenberger and Lachin (2016) interpret the null hypothesis in the randomization tests in the following:

> Under the null, each patient's observed responses is what would have been observed regardless of whether treatment A or B had been assigned. Then the observed difference between the treatment groups depends only on the way in which the n patients were randomized.

A measure of difference between treatments is selected as the test statistic when implementing the randomization test. The test statistic based on the observed randomization sequence is denoted $S_{obs.}$. Under the null hypothesis, we assume there is no difference between treatment A and B. When implementing the randomization test, randomized sequences are generated according to the randomization procedure adopted in practice. Each sequence is generated with a certain probability. The set that contains all possible randomization sequences is called the reference set. Then for each sequence in the reference set, we compute the corresponding test statistic and compare it with the observed test statistic. The p-value of the observed test statistic is calculated by adding the probability of sequences that have test statistics as the least extreme as the observed test statistic. A small p-value indicates that there are differences between treatments.

The reference set contains all possible sequences as long as sequences are generated under the randomized procedure implemented in the trial. Let Ω be the cardinality of the reference set and S be a test statistic of interest. Let L denote a record of the randomization

sequence, and l be its realization. The method to calculates p-values for the randomization tests is described as follows. For a randomization test, the two-sided p-value is given by:

$$p = \sum_{l=1}^{\Omega} l(|S_l| \geq |S_{obs.}|)\Pr(L = l).$$

If the p-value, is less than the given significance level α, we would make the statement that there is a significant difference in treatment effects between treatments A and B.

The flexibility in the selection for the test statistic of interest S makes the randomization test even more versatile under different types of outcomes. For instance, S can be a difference between group means or group proportions. We can also use rank scores. Let a_{jn} be the score of patient j, where n patients are involved in the trial with outcomes $Y_1, \ldots, Y_n$. Let $\bar{a}_n$ be the arithmetic mean of a_{jn} where $j = 1, \ldots, n$. We have a test statistic for the randomization test based on a linear rank test. The test statistic is given by

$$S = \sum_{j=1}^{n} (a_{jn} - \bar{a}_n)T_j,$$

where $T_j = 1$ if patient j received treatment A and $T_j = 0$ if treatment B. If a_{jn} are simple ranks, then the test becomes the Wilcoxon rank-sum test. For binary responses, $a_{jn} = 1$ or 0.

When implementing a randomization test, Monte-Carlo re-randomization is used when generating the possible randomization sequences in the reference set under the given randomization procedure. If enough sequences are generated, replicates will mirror the probability distribution of the corresponding reference set. The two-sided p-value is estimated by

$$\hat{p} = \frac{\sum_{l=1}^{L} I(|S_l| \geq |S_{obs.}|)}{L}.$$

In Galbete and Rosenberger (2016), it was demonstrated by simulation that the test under $L = 15000$ sequences are almost identical to the exact test (where the probability of each sequence in the reference set are computed directly) and even more accurate than asymptotic tests for moderate sample sizes.

The performance of the the randomization test has been tested by simulation. The power of the randomization test is close to the t-test when no time trend exists. When there is a time trend in patients' responses, the randomization test is less affected than the t-test in terms of power. More details about time trends will discuss in the following.

1.4 Missing Data Mechanisms

Missing data can compromise the validity of statistical inference from a randomized clinical trial, especially when the missingness is not handled appropriately. In this case, clarifying the mechanism resulting in the missingness is essential to the analysis. The concepts of missing-data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), are formulated in terms of likelihoods in Little and Rubin (2002). We introduce the missing data mechanisms by using the notation in Little and Rubin (2002).

Let $\boldsymbol{Y} = (y_{ij})$ defined the $(n \times K)$ complete data set with ith row $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{iK})$ where y_{ij} is the value of variable $\boldsymbol{Y_j}$ for observation i. With missing data, define the missing data indicator matrix $\boldsymbol{M} = (m_{ij})$ such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The missing-data mechanisms are presented in terms of conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{Y}$, $f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi})$ where $\boldsymbol{M}$ is parametrized by unknown parameter $\boldsymbol{\psi}$.

1.4.1 Missing Completely at Random (MCAR)

Data are said to be missing completely at random (MCAR) if absence or the presence of observations does not depend on the observed or unobserved data. In terms of the

conditional distribution, the MCAR is described as the following:

$$f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{\psi}) \text{ for all } \boldsymbol{Y}, \boldsymbol{\psi}.$$

1.4.2 Missing at Random (MAR)

Let $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ where $\boldsymbol{Y}_{obs}$ is defined as the observed component of $\boldsymbol{Y}$ and $\boldsymbol{Y}_{mis}$ is the missing component. When the missingness depends on the observed data of $\boldsymbol{Y}$, $\boldsymbol{Y}_{obs}$, the missing-data mechanism is said to be missing at random (MAR) if

$$f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \text{ for all } \boldsymbol{Y}_{mis}, \boldsymbol{\psi}.$$

1.4.3 Missing not at Random (MNAR)

If the distribution of $\boldsymbol{M}$ depends the data in $\boldsymbol{Y}$, the missing-data mechanism is called missing not at random (MNAR). MNAR is also known as nonignorable nonresponse, or not missing at random (NMAR).

1.5 Population-based Methods for Handling Missing Data

The following are frequently-used population-based methods to handle missing data problems; they are the complete-case analysis, single imputation methods, and model-based methods such as the direct likelihood method and multiple imputation. In this section, we will briefly review these methods.

1.5.1 Complete Case Analysis

The complete-case analysis only analyzes the complete observations in the dataset; it is the most commonly used method in practice since it seems to be the quickest and easiest way to solve the missing data problem. It provides statistical inference without bias under the MCAR assumption since the sample under MCAR remains representative for the entire

dataset because the missing does not depend on observed or unobserved data. However, due to the decreased sample size, the complete-case analysis decreases the statistical power of a test and increases the standard errors of an estimator. In practice, it is nearly impossible to ascertain if the missing data are MCAR. Thus simply assuming MCAR and implementing the complete-case analysis potentially risks the statistical inference.

1.5.2 Single Imputation

Imputation for the missing values is conducted once in single imputation methods. Imputed values replace the missing values. After imputation, all observations become complete in the original dataset, and the dataset with imputed values is available to be analyzed. Many single imputation methods are available, and they are reviewed in the following.

- Mean imputation

  This method applies to continuous variables. The missing values are replaced by the mean calculated based on the complete observations in the dataset. Mean imputation is a quick and easy way to fix the missing data problem. However, this method should always be used with caution. As described in Van Buuren (2018), mean imputation underestimates the variance, disturbs the relations between variables, and biases almost any estimates other than the mean when data are not MCAR.

- Regression imputation

  This method applies to continuous variables. First, a regression model is built based on complete observations. Then the missing values will be replaced by the prediction from the regression model built on complete observations. The regression method generates unbiased estimates of the mean under MCAR. However, this method is risky to use because of its potential problems such as underestimating the variance and artificially strengthening the relations in data. Van Buuren (2018) also points out the following:

> Regression imputation, as well as its modern incarnations in machine learning is probably the most dangerous of all methods described here. We may be led to believe that we're to do a good job by preserving the relations between the variables. In reality however, regression imputation artificially strengthens the relations in the data. Correlations are biased upwards. Variability is underestimated. Imputations are too good to be true.

- Stochastic regression imputation

  Similar to the regression imputation method, a model is first built based on complete observations. The stochastic regression method's imputed value is the predicted value plus a random draw from the estimated distribution of residuals from the regression model built on complete observations. Adding noise to the predicted value is an attempt to alleviate the problem that the relations between variables are artificially strengthened because of imputation.

- Predictive mean matching (PMM)

  Instead of imputing missing values by a predetermined model, the predictive mean matching method selects observed values from complete cases to replace missing values. Van Buuren (2018) described the predictive mean matching method:

  > For each missing entry, the method forms a small set of candidate donors typically with 3, 5, or 10 members from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor is randomly drawn from the candidates, and the observed value of the donor is taken to replace the missing value.

  There are several methods to select donors based on metrics defined to measure the similarity between the predicted values for the missing entry and the predicted values from complete cases. Methods for selecting donors are detailed in Section 3.4.2 of Van Buuren (2018). Some discussion on how to decide the number of donors is briefly introduced in Section 3.4.3 of Van Buuren (2018). In general, the number of donors

depends on the sample size. Setting the number of donors equal to 3 or 5 is usually used in practice.

The advantage of applying PMM is that it will not impute any implausible values outside the range of the observed values. Thus, even when the model used to impute missing values is misspecified, the PMM is less vulnerable than previously discussed methods (e.g., the regression imputation method). The PMM method is valid for both continuous and binary missing values. It should be noted that it will be more reasonable to use the PMM method for continuous missing values than for binary values if the model used to predict missing values is a linear regression model.

- Logistic regression imputation

  This method applies to binary variables. It is implemented through Bayesian logistic regression. First, logistic regression is fit based on complete data and coefficients $\hat{\boldsymbol{\beta}}$ and variance of coefficients $\boldsymbol{V} = var(\hat{\boldsymbol{\beta}})$ are estimated via iteratively reweighted least squares. Then a $\dot{\boldsymbol{\beta}}$ is drawn from the multivariate normal distribution, which is built based on $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{V}$; see Van Buuren (2018) Section 3.6.1 for the exact procedure of drawing $\dot{\boldsymbol{\beta}}$. Then the predicted probability $\dot{p}$ for each missing response $\dot{y}$ with its predictor $\dot{\boldsymbol{X}}$ is calculated; i.e., $\dot{p} = 1/\left(1 + exp(-\dot{\boldsymbol{X}}\dot{\boldsymbol{\beta}})\right)$. A random variable U is generated from the uniform distribution $U(0, 1)$. Then, imputation for each missing response $\dot{y}$ is calculated, where $\dot{y} = 1$ if $u \leq \dot{p}$ and $\dot{y} = 0$ otherwise. This is implemented using the logreg() function in the R package "**Mice**".

  However, "perfect prediction" may occur in practice. White et al. (2010) discussed this perfect prediction problem and its potential harm:

  > Perfect prediction may occur in any GLM with a categorical outcome. In this case, the likelihood tends to a limit as one or more regression parameters go to plus or minus infinity: loosely, these parameters have maximum likelihood estimate (MLE) equal to plus or minus infinity. It is arguable

whether this is in itself a problem, since odds ratios of 0 or infinity should be no more surprising than estimated probabilities of zero or one. However, a problem definitely arises with standard errors computed from the information matrix: these are extremely large, reflecting the near-flat nature of the likelihood.

The exact solution to the perfect prediction problem depends on the choice of software. Details are how to implement the method is also available in Section 3.6.2 of Van Buuren (2018).

- "Worst-rank" method

  This method was proposed in Lachin (1999), and it applies to both binary and continuous variables. Lachin describes a worst-rank analysis, i.e., assigning more extreme values (values indicating "worst" treatment effects) than observed values as the imputed values for missing data. All missing values share the same values (ranks) if a worst-rank analysis applies.

- "Best-worst and worst-best" method

  We call it "Best - worst" method in the following sections. This method works for both binary and continuous variables. Suppose we have missing values in patients' responses if a two-armed randomized clinical trial is implemented. The experimental group (treatment A) tends to have more beneficial outcomes. The control treatment group (treatment B) tends to have less beneficial responses. If the "best - worst" method is adopted, the imputed values for missing responses in treatment A will represent harmful outcomes (i.e., the "worst" values among observed values). While in treatment B, the imputed values will represent beneficial outcomes (i.e., the "best" values among observed values).

1.5.3 Direct Maximum Likelihood Method

The direct maximum likelihood method, sometimes called "full information maximum likelihood" or just "maximum likelihood," is a method for handling missing data without imputing missing values under MAR. We will review the direct maximum likelihood method by following the notation in Section 6.2 of Little and Rubin (2002). Let $\boldsymbol{Y}$ define the data with missing values where $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$, $\boldsymbol{Y}_{obs}$ denotes the observed value and $\boldsymbol{Y}_{mis}$ denotes the missing value; and $\boldsymbol{Y}$ is parametrized by some parameter $\boldsymbol{\theta}$ that we want to estimate. Define $f(\boldsymbol{Y}|\boldsymbol{\theta}) \equiv f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta})$ as the density of joint distribution of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$, and the marginal probability density of $\boldsymbol{Y}_{obs}$ is

$$f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}) = \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}) d\boldsymbol{Y}_{mis}.$$

The likelihood of $\boldsymbol{\theta}$ based on data $\boldsymbol{Y}_{obs}$ ignoring the missing-data mechanism (denoted as L_{ign}) is a function of $\boldsymbol{\theta}$ proportional to $f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta})$ where

$$L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \propto f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}.$$

If the missing data mechanism is ignored, inferences about $\boldsymbol{\theta}$ can be derived from the likelihood $L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$. As before, let $\boldsymbol{M}$ define the missing data indicator of $\boldsymbol{Y}$. $\boldsymbol{M}$ is treated as a random variable. The joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{M}$ is given by:

$$f(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{Y}|\boldsymbol{\theta}) f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi}), \quad (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Omega_{\boldsymbol{\theta},\boldsymbol{\psi}},$$

where $\Omega_{\boldsymbol{\theta},\boldsymbol{\psi}}$ is the parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$, and the conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{Y}$ is indexed by unknown parameter $\boldsymbol{\psi}$. In practice, the actual observed data contain the

value of variables $(\boldsymbol{Y}_{obs}, \boldsymbol{M})$. The distribution of the observed data is given by:

$$\begin{aligned} f(\boldsymbol{Y}_{obs}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{Y}_{mis} \\ &= \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}) f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) d\boldsymbol{Y}_{mis} \end{aligned}$$

The full likelihood of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is any function of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, and it is given by:

$$L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{M}) \propto f(\boldsymbol{Y}_{obs}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}), \quad (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Omega_{\boldsymbol{\theta},\boldsymbol{\psi}}.$$

Notice that when missing-data mechanism is MAR,

$$f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \text{ for all } \boldsymbol{Y}_{mis}.$$

The distribution of $\boldsymbol{Y}_{obs}, M$ becomes

$$f(\boldsymbol{Y}_{obs}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \times \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta}) d\boldsymbol{Y}_{mis} = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}).$$

As Little and Rubin (2002) point out:

> In many important practical applications, the parameter $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, in the sense that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\psi})$, is the product of the parameter space of $\boldsymbol{\theta}$ and the parameter space of $\boldsymbol{\psi}$, $\Omega_{\boldsymbol{\theta},\boldsymbol{\psi}} = \Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\psi}}$. If the mechanism is MAR and $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are distinct, then likelihood-based inferences for $\boldsymbol{\theta}$ from $L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{M})$ will be the same as likelihood-based inference for $\boldsymbol{\theta}$ from $L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$, since the resulting likelihoods are proportional.

Moreover, the direct maximum likelihood has been shown to provide unbiased parameter estimates and standard errors under MAR and MCAR by Enders and Bandalos (2001).

Consider the linear regression model with the following format

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}^{\intercal}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\tilde{\boldsymbol{X}} = (\mathbf{1}, \boldsymbol{X}^{\intercal})^{\intercal}$ and the predictors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^{\intercal} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, and the $\epsilon_i \sim N(0, \sigma^2)$. Some values are missing within n independent observations $(y_i, \boldsymbol{X}_i^{\intercal})^{\intercal}$ where $i = 1, \ldots, n$. Based on the normality and independent assumptions of $\boldsymbol{y}$ and $\boldsymbol{X}$, we have

$$(\boldsymbol{y}, \boldsymbol{X}) \sim N(\mu_{\boldsymbol{y},\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{X}})$$

where $\boldsymbol{\mu}_{\boldsymbol{y},\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{y}} \\ \boldsymbol{\mu}_{\boldsymbol{X}} \end{pmatrix}$ and $\boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{y}} & \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{X}} \\ \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{y}} & \boldsymbol{\Sigma}_{\boldsymbol{X}} \end{pmatrix}$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_{\boldsymbol{y},\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{X}})$ are estimated through the expectation maximization algorithm (EM) algorithm and it is implemented by the R package **NORM** (Novo and Schafer (2013)). Details are available in Section 5.3 of Schafer (1997).

Implementation of the direct maximum likelihood method via the EM algorithm is briefly discussed in the context of the linear regression with missing values. We are interested in estimating the unknown parameter $\boldsymbol{\theta}$. The data $\boldsymbol{Y}$ is defined as in the previous section, where $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$. Under the MAR assumption, likelihood-based inferences for $\boldsymbol{\theta}$ from $L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}_{obs}, M)$ will be the same as likelihood-based inferences for $\boldsymbol{\theta}$ from $L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$, i.e. maximizing the observed data log-likelihood which is given by:

$$l(\boldsymbol{\theta}; \boldsymbol{Y}_{obs}) = \log f(\boldsymbol{Y}_{obs}; \boldsymbol{\theta}) = \log \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\theta}) d\boldsymbol{Y}_{mis}.$$

However, this maximization is difficult to achieve in practice. We can achieve the MLE of $\boldsymbol{\theta}$ by iteratively maximizing the expected complete data log-likelihood which is given by:

$$l(\boldsymbol{Y}; \boldsymbol{\theta}) = \log f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\theta}).$$

Steps of EM algorithm are listed in the below:

- Initiate $\boldsymbol{\theta}^{(0)}$; $\boldsymbol{\theta}^{(t)}$ is the estimate of $\boldsymbol{\theta}$ at the tth iteration.
- E step: compute the expectation of complete-data log-likelihood with respect to the conditional distribution of $\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}$ with $\boldsymbol{\theta}^{(t)}$, i.e.:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[l(\boldsymbol{Y};\boldsymbol{\theta})|\boldsymbol{Y}_{obs};\boldsymbol{\theta}^{(t)}] = \int l(Y;\boldsymbol{\theta}) f(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs};\boldsymbol{\theta}^{(t)}) d\boldsymbol{Y}_{mis}.$$

- M step: maximize the Q function to obtain $\boldsymbol{\theta}^{(t+1)}$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

- Iterate between E step and M step until the change in function Q is very small.
- The estimate of $\boldsymbol{\theta}$ based EM algorithm is obtained.

After $\boldsymbol{\theta}$ has been estimated by EM algorithm, we can have the estimate of the coefficient $\boldsymbol{\beta}$ through the following. With

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\mu_y} - \boldsymbol{\Sigma_{y,X}}\boldsymbol{\Sigma_X^{-1}}\boldsymbol{\mu_X} + \boldsymbol{\Sigma_{y,X}}\boldsymbol{\Sigma_X^{-1}}\boldsymbol{X},$$

then the coefficient $\boldsymbol{\beta}$ was estimated by the following form

$$\boldsymbol{\beta} = (\boldsymbol{\mu_y} - \boldsymbol{\Sigma_{y,X}}\boldsymbol{\Sigma_X^{-1}}\boldsymbol{\mu_X}, \boldsymbol{\Sigma_{y,X}}\boldsymbol{\Sigma_X^{-1}})^{\intercal}.$$

The standard deviations can be estimated by the following forms

$$\mathbb{V}[\boldsymbol{\beta}] = \mathrm{diag}(\boldsymbol{C}), \quad \text{with}$$
$$\boldsymbol{C} = \left(\boldsymbol{\Sigma_y} - \boldsymbol{\beta}^{\top}\boldsymbol{\Sigma_X}\boldsymbol{\beta}\right)\left(\left(\mathbf{0}_{p+1}, (\mathbf{0}_p, \boldsymbol{\Sigma_X})^{\top}\right)^{\top} + \left(\mathbf{1}, \boldsymbol{\mu_X}^{\top}\right)^{\top}\left(\mathbf{1}, \boldsymbol{\mu_X}^{\top}\right)\right)^{-1}/n$$

After obtaining the estimated $\boldsymbol{\beta}$ and its standard errors, a t-test can be conducted to test the significance of a predictor's coefficient.

1.5.4 Multiple Imputation

Multiple imputation (MI) is an alternative method for dealing with missing data under MAR.In general, MI contains three steps:

- **Imputation step**

  In this step, multiple imputations are conducted. Hence, multiple completed datasets are generated by replacing the missing values with the imputed values multiple times. Usually, obtaining 50 or more than 50 imputations is acceptable to reduce the sampling uncertainty from the imputation process.

- **The complete-data analysis step**

  A desirable statistical analysis is conducted individually on each complete dataset generated from the previous step.

- **Pooling step**

  Collect statistical inference results (e.g., parameter estimates and their standard errors) from the previous step. Rubin (2004) proposed a set of rules for combining the separate estimates and standard errors from each of the imputed dataset into an overall estimate with standard error, confidence intervals, and p-values. These rules are based on asymptotic theory on the normal distribution and are implemented in the functions pool() and pool.scalar() via the R package "**Mice**" (Van Buuren and Groothuis-Oudshoorn (2010)).

1.5.5 Which Method We Should be Adopted under Different Missing Data Mechanisms?

In general, the least stringent assumption is MCAR if we wish to apply the methods discussed above. Under MCAR, the complete-case analysis is valid to handle missing data,

and it is only valid under MCAR. When assuming MAR, the direct maximum likelihood method and multiple imputation are options to handle the missingness. Single imputation methods such as the mean imputation, the regression imputation should always use with caution due to their potential problems (detailed discussion is available in Section 1.5.2). Van Buuren points out the case when the complete-case analysis and multiple imputation are equivalent in Section 2.7 of Van Buuren (2018):

> Suppose that the complete-data model is a regression with outcome Y and predictors X. If the missing data occur in Y only, complete-case analysis and multiple imputation are equivalent, so then complete-case analysis is preferred since it is easier, more efficient and more robust by Von Hippel (2007).

None of the methods we discussed so far can deal with MNAR. Particular assumptions about the missing mechanism are required to proceed with data analyses under MNAR, and the related discussion is beyond the scope. In practice, it is nearly impossible to ascertain the missing data is MAR or MCAR because no information is provided about the missing values. As Schafer and Graham (2002) point out:

> When missingness is beyond the researcher's control, its distribution is unknown, and MAR is only an assumption. In general, there is no way to test whether MAR holds in a dataset, except by obtaining follow-up data from nonrespondents or by imposing an unverifiable model.

1.6 Outline of the Thesis

Chapter 2 reviews and proposes randomization-based missing data methods: the unconditional reference set, the conditional reference set, and randomization-based multiple imputation. Methods discussed in Chapter 2 deal with the case when missing values exist in continuous/binary patients' responses when a two-armed randomized clinical trial is implemented. Which method should be selected in terms of missing data mechanisms is discussed in Chapter 2. Chapter 3 describes how different simulation scenarios are set

up to compare randomization-based missing data methods and population-based methods. The simulation assumes a two-armed clinical trial with complete treatment assignments and continuous/binary responses with missing values. Details about software implementation in simulation studies are also presented in Chapter 3. Chapter 4 contrasts the parametric missing data inference with randomization-based methods without heterogeneity in simulated patients' responses. The contrast is conducted based on the simulation scenarios proposed in Chapter 3. Methods comparisons are conducted separately for continuous and binary patient responses. In Chapter 5, parametric and randomization-based missing data inference are contrasted when assuming heterogeneity existed in continuous patient responses. In Chapters 4 and 5, the comparison is conducted via simulated type I error rates and statistical power of parametric and randomization-based missing data methods. Chapter 6 proposes a randomization-based confidence interval to determine the missing data mechanism. Finally, future work and concluding remarks of parametric and randomization-based missing data inference for randomized clinical trials are presented in Chapter 7.

Chapter 2: Methods for Randomization-Based Inference with Missing Data

2.1 Randomization Tests with Missing Data

Randomization tests provide a natural way to analyze data from randomized clinical trials since no random sampling assumption is required. Randomization tests can be used to analyze nearly all types of primary outcomes encountered, including continuous, binary, ordinal, survival with censoring, rates of change from longitudinal models, and adjusted treatment effects from regression models with covariates.

While there is a rich literature on parametric methods to handle missing data, no available analysis exists for randomization tests. It is naive and risky to adopt the complete-case analysis since it is only valid under MCAR. Also, excluding observations with missing values, especially in randomized clinical trials, is anathema for people implementing the randomization test since the distribution of the randomization sequence is not preserved.

This section formalizes a randomization test when the missing data problem occurs during a randomized clinical trial. Randomization-based methods handling missing data are described in terms of two types of reference sets:

- The unconditional reference set where the missing values do not impact the test statistic's computation.
- The conditional reference set that conditions on the missing data pattern and the number of missing values on each treatment.

Both methods are proposed but not developed or formalized in Edgington and Onghena (2007) Section 14.16, and reference sets contain equiprobable randomization sequences in their context. In Kennes et al. (2012), some properties of two references sets were explored

only under the RAR with small sample sizes, and no global recommendations were made. We discuss randomized-based inference under the framework of a linear-rank test. Linear rank tests form a large family of tests that incorporates continuous, binary, ordinal, and time-to-event outcomes, as well as covariate-adjusted analyses.

To illustrate randomization-based methods, a corresponding probability measure $\mathcal{P}(\boldsymbol{T})$ is defined. The reference set $\Omega = \{\boldsymbol{t} \in \{0,1\}^n : \mathcal{P}(\boldsymbol{T} = \boldsymbol{t}) > 0\}$ is formed with all possible realization of the random vector $\boldsymbol{T}$ (i.e., treatment assignments). In the context of a two-armed randomized clinical trial with experimental treatment A and placebo (treatment B), missing values only occur in patients' responses. If patient j receives treatment A then $t_j = 1$ otherwise $t_j = 0$ if treatment B. Define the number of allocations to the treatment A as $N_A = \sum_{j=1}^{n} t_j$ with realization n_A if n patients are in a trial. The missing pattern in responses is defined by vector $\boldsymbol{M}$. Note that $\mathbf{1} \notin \boldsymbol{M} = (m_1, ...m_n)^\intercal \in \boldsymbol{M} = \{0,1\}^n$ and

$$m_j = 1 \text{ if patient } j\text{'s response is missing,}$$

$$= 0 \text{ if patient } j\text{'s response is observed.}$$

The total number of complete observations is $m_\bullet = n - \sum_{j=1}^{n} m_j$ and $n - m_\bullet$ is the total number of missing observations.

2.2 The Unconditional Reference Set

The implementation of a randomization test requires a reference set. The unconditional reference set is adopted to handle missing values, based on the assumption that the missingness is independent of treatment assignments, i.e., $\boldsymbol{M}$ is independent of $\boldsymbol{T}$. This assumption is analogous, in some sense, to the idea of MCAR while we are not discussing the missing data mechanism in terms of the likelihood since the distribution of the test statistic is determined by the reference set, not the likelihood. A $\boldsymbol{\chi^2}$-test may be able to identify deviations from stochastic independence if there are a large number of missing values, but the test would

have very little power for minimal amounts of missing data. For a small amount of missing, the Fisher's exact test can serve as an alternative. Randomization-based confidence intervals are proposed to measure the deviation of the independence assumption between missingness in responses and the treatment assignments by treating missing or not missing in the observed data as binary outcomes. Details about the proposed methodology are available in Chapter 6.

Suppose the linear rank test statistic is adopted for a randomization test. Define $\boldsymbol{a}$ as the vector of non-missing observations' centered ranks where $\boldsymbol{a} = (a_1 - \bar{a_\bullet}, ..., a_{m_\bullet} - \bar{a_\bullet})^\intercal$ and $\bar{a_\bullet} = \frac{1}{m_\bullet}\sum_{j=1}^{m_\bullet} a_j$. The test statistic is given by

$$S_m(\boldsymbol{t}) = \boldsymbol{a}^\intercal \boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}$$

where $\boldsymbol{A}(\boldsymbol{m})$ is a $(m_\bullet \times n)$ matrix with 0 or 1 resulting from the unit matrix by deletion of $(n - m_\bullet)$ rows corresponding to the entries of the $\boldsymbol{m}$ vectors. This test-statistic is the same as the test statistic of a linear rank test when no missing observations; see Rosenberger and Lachin (2016) for details. The original reference set Ω (when all observations are complete) in a randomization test has been reduced to the reference set $\tilde{\Omega}(\boldsymbol{m})$ with randomization sequences $\tilde{\boldsymbol{t}} = \boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}$ and the corresponding probability measure of $\tilde{\boldsymbol{T}}$ is

$$\mathcal{P}(\tilde{\boldsymbol{T}} = \tilde{\boldsymbol{t}}) = \sum_{\boldsymbol{t}:\ \tilde{\boldsymbol{t}} = \boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}} \mathcal{P}(\boldsymbol{T} = \boldsymbol{t}).$$

Then the two-sided p-value for the test statistic becomes

$$p_{\boldsymbol{u}} = I(|S(\tilde{\boldsymbol{t}})| \geq |S_{obs.}|) \sum_{\boldsymbol{t}:\ \tilde{\boldsymbol{t}} = \boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}} \mathcal{P}(\boldsymbol{T} = \boldsymbol{t}).$$

where $S_{obs.}$ is the observed test statistic calculated from the observed randomization sequence. In Kennes et al. (2012), the reference set $\tilde{\Omega}(\boldsymbol{m})$ is called the unconditional reference

set. The estimation of the two-sided p-value $p_{\boldsymbol{u}}$ corresponding to the test statistic S_m is calculated by the Monte-Carlo method. The calculation of the p-value can be implemented by imputing missing responses' centered ranks as zero in the linear rank test and keeping the original reference set Ω. A p-value less than α indicates two treatment effects differ.

The following example gives a further explanation. Suppose we have 5 observations with treatment assignment $\boldsymbol{t}_{obs} = (1, 0, 1, 0, 0)^\intercal$. The 3rd and 4th patients' responses are missing. Then

$$\boldsymbol{A}(\boldsymbol{m}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and $\boldsymbol{A}(\boldsymbol{m})\boldsymbol{t} = (1, 0, 0)$. From the vector $\boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}$, it is clear that patients with missing responses does not contribute to the calculation of the test statistic $S_m(\boldsymbol{t})$. To be more specific, suppose we sample randomization sequences such as $\boldsymbol{t}' = (1, 0, 0, 0, 0)^\intercal$ or $\boldsymbol{t}'' = (1, 0, 0, 1, 0)^\intercal$ or $\boldsymbol{t}''' = (1, 0, 1, 1, 0)^\intercal$, the resulting $\boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}$ remains to be $(1, 0, 0)$ when implementing a randomization test. No matter how the treatments are allocated within the missing data positions (3rd and 4th patients), it does not affect the calculation of the test statistic. Implementing this procedure is based on the assumption that missingness does not depend on the treatment assignments. In other words, the missing observations themselves and the proportion of treatment assignments within missing data positions do not affect the calculation of the test statistic.

If the differences in group means is adopted as the test statistic, the mean value calculated from all complete values is imputed as the replacement for missing values. After missing values are imputed, a randomization test is applied.

2.3 The Conditional Reference Set

We assume missingness does not depend on treatment assignments when adopting the unconditional reference set in a randomization test. However, this assumption is not always

valid. We may observe a trend that the proportion of missing observations on each treatment differs substantially, which indicates the dependent relationship between treatment assignments and the missingness. The missingness depending on treatments, in some sense, is analogous to MAR. If missingness depends on the treatment assignments, the conditional reference set is adopted to handle the missingness. The conditional reference set only contains randomization sequences that have same missing data positions and the same number of missing on each treatment as the observed randomization sequence. More formally, denote the random variable $M_A = \boldsymbol{T}^\intercal \boldsymbol{m}$ as the number of missing responses in the experimental treatment A, with realization m_A calculated from the observed randomization sequence. Then the two-sided p-value conditioning on $\boldsymbol{M} = \boldsymbol{m}$ and $M_A = m_A$ is

$$\begin{aligned} p_c &= \sum_{\boldsymbol{t}\in\Omega} I(|S_{\boldsymbol{m}}(\boldsymbol{t})| \geq |s_{obs}|) P(\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{M} = \boldsymbol{m}) \\ &= \sum_{\tilde{\boldsymbol{t}}\in\tilde{\Omega}(\boldsymbol{m})} \sum_{\tilde{\boldsymbol{t}}=\boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}} I(|S(\tilde{\boldsymbol{t}})| \geq |S_{obs.}|) P(\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{M} = \boldsymbol{m}) \\ &= \sum_{\tilde{\boldsymbol{t}}\in\tilde{\Omega}(\boldsymbol{m})} I(|S(\tilde{\boldsymbol{t}})| \geq |S_{obs.}|) \sum_{\tilde{\boldsymbol{t}}=\boldsymbol{A}(\boldsymbol{m})\boldsymbol{t}} P(\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{M} = \boldsymbol{m}). \end{aligned}$$

The estimation of two-sided p-value p_c corresponded to test statistic $\boldsymbol{S_m}$ is also calculated based on a Monte-Carlo method. For the conditional reference set, we first generate sequences unconditionally and then only keep the sequences that have the same number of missing observations on each treatment as the observed sequence. The missing positions are the same as the observed ones. Sequences that do not satisfy these conditions are eliminated. The probabilities are reweighted accordingly within the conditional reference set. It is analogous to the idea of using weighted generalized estimating equations to reweight components of the likelihood to account for the probability of missingness; see Molenberghs and Verbeke (2006) for details.

For the conditional reference set method, the way to impute missing values is the same as the one for the unconditional method. To be more specific, if the difference in group means

is adopted as the test statistic for a randomization test, then the imputed value would be the mean from all complete values (patients' responses); if the linear-rank test statistic is adopted, the imputed centered rank (for a patient's response) would be zero. The difference is that the conditional reference set uses a different reference set - the reference set that only contains sequences with same number of missing values on each treatment, on the locations having missingness.

The conditional method and the unconditional method maintain the spirit of the randomization test, and these are two practicable alternatives in handling missing data when a randomization test is applied. Simulation studies are conducted to investigate the performance of these two methods.

2.4 Randomization-Based Multiple Imputation (RBMI)

Randomization-based methods based on of the unconditional or conditional reference set are analogous to the single-imputation in some sense since the imputed rank for each missing response is the same; i.e., all imputed centered ranks for missing responses are zero. When handling missing values, not incorporating the uncertainty in data imputation makes single imputation questionable, especially when we are less confident about the imputation method. In contrast, multiple imputation provides a better solution to covering the uncertainty from data imputation. The merit of multiple imputation has been pointed out by Van Buuren (2018):

> Multiple imputation is unique in the sense that it provides a mechanism for dealing with the inherent uncertainty of the imputations themselves.

To improve single imputation, we propose a randomization-based multiple imputation. The algorithm of a randomization-based multiple imputation is sketched in Figure 2.1. Suppose we obtain complete treatment assignments and patients' responses with missing values in a two-armed clinical trial. Based on the observed data, N complete datasets are imputed by methods that can introduce the uncertainty in data imputation, such as the PMM method.

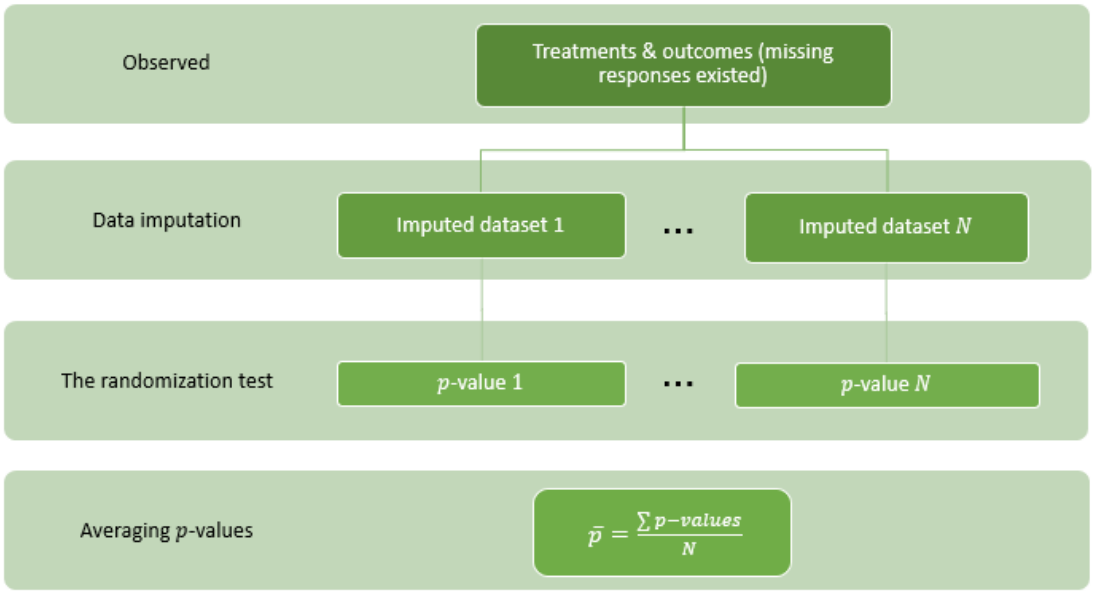

Figure 2.1: *Algorithm of the randomization-based multiple imputation*

Once a complete dataset is imputed, an unconditional randomization test is applied; and a p-value is obtained. Multiple p-values are collected and averaged next. Finally, the averaged p-value, $\bar{p}$, is compared with the adopted significance level. If $\bar{p}$ is less than the significance level, the null hypothesis (assuming no difference in treatment effects between groups) is rejected.

Chapter 3: Simulation Protocol

The performance of randomization-based missing data methods is investigated by comparing them with other population-based missing data methods. Especially, the performance of randomization-based methods is of interest under homogeneous or heterogeneous patient responses. Multiple simulation scenarios are simulated, and randomization-based methods are compared with population-based methods in terms of type I error rates and power, based on 1000 replicates for each simulation case. Simulation under homogeneous responses is discussed in Chapter 4, and simulation under heterogeneous responses is presented in Chapter 5.

3.1 Simulation Setting

We introduced several population-based and randomization-based methods to handle missing data in the Section 1.5 and Chapter 2, respectively. In the context of a two-armed clinical trial with treatment A and B, we are interested in investigating these methods' performance when assuming equality and inequality in treatment effects between A and B in the null hypothesis. Assume there are n observations (patients) in the dataset (trial), and $n = 50$ or 100 in simulation studies. Treatment assignments are generated based on a randomization procedure that we discussed in previous sections; options for procedures are CR, RAR, PBD (blocksize = 4 or 6), TBD, RBD (maximum blocksize = 6), BSD. Two types of responses are generated. Binary responses are simulated from Bernoulli distributions; continuous responses are from normal distributions. When responses are continuous, higher values in responses represent better treatment effects. When responses are binary, the response equals one representing a success. Before we simulate the missingness in responses, complete responses are generated based on the parameters in the Table 3.1 below.

Table 3.1: *Homogeneous responses: distributions for simulating responses*

	Type I error rate	Power, $n = 50$	Power, $n = 100$
Binary	A: Ber(0.2)	A: Ber(0.65)	A: Ber(0.5)
	B: Ber(0.2)	B: Ber(0.2)	B: Ber(0.2)
Continuous	A: N(0.2,1)	A: N(1.2,1)	A: N(0.9,1)
	B: N(0.2,1)	B: N(0.2,1)	B: N(0.2,1)

Table 3.2: *Simulate missingness in binary responses based on missing mechanisms*

Missing data mechanism	How to simulate missing values
MCAR	The probability that a patient's response is missing is p (i.e., Ber(p_{ms}), $p_{ms} = 0.05, 0.1$).
MAR	$p_B : p_A = r$; $r = 1 : 3$
MNAR	$p_f : p_s = r$; $r = 1 : 3$

After treatment assignments and responses are simulated, the missingness in responses is generated according to missing data mechanisms: MCAR, MAR, and MNAR. The overall proportion of missing in the responses is controlled by the probability p_{ms}. The details about how to simulate missing responses are in Table 3.2 and 3.3 below. No missing values occur in the treatment assignments.

- N_A: the total number of patients receiving treatment A;
- N_B: the total number of patients receiving treatment B;
- p_A: the missing probability if the patient receives treatment A;
- p_B: the missing probability if the patient receives treatment B;

Table 3.3: *Simulate missingness in continuous responses based on missing mechanisms*

Missing data mechanism	How to simulate missing values
MCAR	The probability that a patient's response is missing is p_{ms} (i.e., Ber(p_{ms}), $p_{ms} = 0.05, 0.1$).
MAR	$p_B : p_A = r$; $r = 1 : 3$
MNAR	Right - tailed logistic distribution function*

*: Implemented by ampute in R package: **mice**. Larger values in responses have higher probabilities of being missing. The logistic function is shifted according to the overall missing proportion p_{ms}. Details are available in the following section

- N_s: the total number of success;
- N_f: the total number of failures;
- p_s: the missing probability if the outcome is a success;
- p_f: the missing probability if the outcome is a failure.
- Parameters N_A, N_B, N_s, N_f are counts calculated on the simulated complete dataset.
- The exact value for p_A, p_B, p_s, p_f in an simulated dataset are decided by the following two equations:

$$N_A \cdot p_A + N_B \cdot p_B = (N_A + N_B) \cdot p_{ms},$$

$$N_s \cdot p_s + N_f \cdot p_f = (N_s + N_f) \cdot p_{ms},$$

  while $p_B : p_A = 1/3$ and $p_f : p_s = 1/3$.

- Bernoulli random variables are used to decide all missing values except for continuous responses under MNAR. For example, when responses are binary, whether the ith patient's response is missing or not is decided by p_s or p_f. If ith patient response is

a success, we use a Bernoulli random variable with probability p_s to decide whether the response is missing or not.

For each missing method studied, given a type of response, multiple simulation cases generated. For Chapter 4 and Section 5.1, several randomization procedures are CR, RAR, BCD, BSD, PBD (blocksize equals 4 or 6), TBD and RBD (maximum blocksize = 6) with sample sizes $n = 50$ or 100, and the overall missing proportion ($p_{ms} = 0.05$ or 0.1). For Section 5.2 and 5.3, the procedures involved are BCD and RBD (maximum blocksize = 6) and sample size $n = 200$. For each simulation case, the type I error rate or power is calculated based on 1000 replications.

3.2 Software Implementation in Simulation Studies

3.2.1 Simulate Missingness in Continuous Responses under MNAR

We utilize the function **ampute** function in R package "**Mice**" to simulate missing values in continuous responses under MNAR. The "**Mice**" package is developed by Van Buuren and Groothuis-Oudshoorn (2010). The **ampute** function serves as a tool for simulating missing values in a multivariate dataset (dataset with multiple columns) while controlling the overall proportion of missing values p_{ms} (noted as prop in the Figure 3.1). Figure 3.1 sketches the procedure of simulating missing values by **ampute** function.

The multivariate amputation procedure of **ampute** is built upon an initial idea proposed in Brand (1999). Details about how to use **ampute** function is available in Schouten et al. (2018).

We target to simulate missing values within a single column (i.e., the vector for patient responses) instead of multiple columns. A small trick is used in order to use **ampute** since it only can be used to simulate missing values in a dataset with multiple columns. Before simulating missing values in continuous responses under MNAR, a dataset containing two columns is simulated by replicating simulated complete patient responses twice. Suppose complete responses are $\boldsymbol{y}_1$, we use $\boldsymbol{y}_1$ as the first column of a complete dataset. We replicate

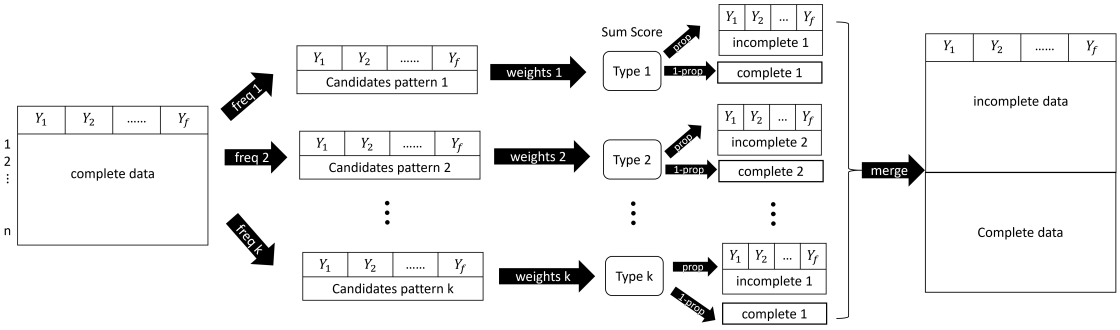

Figure 3.1: *Schematic overview of the multivariate imputation procedure (Schouten et al. (2018), page 2914.)*

the $\boldsymbol{y}_1$ and named it $\boldsymbol{y}_2$, then use $\boldsymbol{y}_2$ as the second column of a complete dataset. Then responses with missing values will be simulated based on the complete dataset.

Some parameters are also needed to be set up in order to use the **ampute**. We assume two missing patterns (i.e., $k = 2$ in the Figure 3.1) for the previous simulated complete dataset.

- Missing pattern 1: we have missing values in $\boldsymbol{y}_1$ and no missing values in $\boldsymbol{y}_2$;
- Missing pattern 2: No missing values in $\boldsymbol{y}_1$ but we have missing values in $\boldsymbol{y}_2$.

After we decide on missing patterns, a complete dataset is randomly divided into two subsets. The frequency of missing patterns decides the number of observations for each subset. We set both frequencies equal to 0.5, which means half of the complete dataset becomes the candidate for missing data pattern one, and the other half is the candidate for missing data pattern two. For each observation in a complete dataset, we have its simulated $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, and the probability of having a missing value in $\boldsymbol{y}_1$ or $\boldsymbol{y}_2$ is controlled by the weighted sum scores (WSS). WSS is calculated as the outcomes of a linear combination

of values from each column. For observation i, WSS is calculated as follows. Suppose f columns (variables), $\boldsymbol{y}_1, ..., \boldsymbol{y}_f$ exist in a complete dataset.

$$wss_i = w_1 \cdot y_{1i} + w_2 \cdot y_{2i} + ... + w_f \cdot y_{fi},$$

where wss_i is defined as the wss for ith observation. Based on our simulation target, we set $w_1 = 1$ and $w_2 = 0$ under missing pattern 1; $w_1 = 0$ and $w_2 = 1$ under missing pattern 2. Since two columns in a complete dataset are the same, only the value from one column affects the calculation of WSS for ith observation (patient). After calculating the WSS for ith patient, the right-tailed logistic distribution function is used. Thus, candidates with high weighted sum scores will receive a high probability of missing. Noted that the logistic distribution function has been adjusted according to patient response and the overall proportion of missing values p_{ms}.

3.2.2 Programming Implementation with Rcpp, Parallel Computing and High Performance Computing

In the context of this project, sometimes we may simulate some observed randomization sequences that have very imbalanced missing values between two treatments. Even though the probability of obtaining such sequences is very small in reality, it is still likely to happen due to the randomness in simulation studies. For example, suppose we have ten missing values in observed patient response; only one missing value is in treatment A, and the remaining nine are all in treatment B. Suppose the conditional reference set method is adopted. The probability of simulating a randomization sequence with one missing value in treatment A and nine in treatment B as the observed sequence is low, making the conditional reference set method take a much longer time finding enough sequences to proceed with a randomization test.

Thus, instead of coding all methods by using R, the implementation of randomization-based missing data inference is coded under the **Rcpp** package in the following simulation

studies. **Rcpp** is a tool developed by Eddelbuettel et al. (2021), that facilitates connecting C++ to R, since C++ runs much faster than base R codes. Since **Rcpp** is implemented under R, combing **Rcpp** and R together in simulation studies can utilize the versatility and flexibility of conducting statistical analysis in R while maintaining the high computation speed of C++.

Parallel computing with the R package "**doParallel**" developed by Calaway et al. (2015) is also applied in the following simulation studies. Parallel computing can utilize multiple CPUs simultaneously to finish the computation workload for one program, which further speeds up simulation studies.

All simulation work is conducted on the high-performance computing clusters-ARGO and HOPPER-with support from the Office of Research Computing at George Mason University.

3.3 Simulation error upper bound for α

Based on γ replication in a simulation, under the significance level $\alpha = 0.05$, the upper bound for a type I error rate is

$$\alpha + 1.96 * \sqrt{\frac{\alpha(1-\alpha)}{\gamma}}$$

where $\alpha = 0.05$. Based on 1000 replications, the upper bound of a type I error rate is about 0.0635 under significance level 0.05. Thus, any simulated type I error rate greater than 0.0635 is considered type I error rate inflation in the following simulation studies.

Chapter 4: Results Under Homogeneity

In this chapter, the performance of randomization-based methods is evaluated by comparing them with population-based methods when patients' responses are homogeneous. Both continuous and binary patients' responses are simulated. Multiple missing data methods are investigated in terms of type I error rates and power. Different methods are adopted under different types of responses.

4.1 Missing Data Methods

The simulated data contains two variables: treatment assignments and patients' responses; missing values occur only in responses. Missing data methods are listed in Table 4.1. Methods' comparisons are conducted separately based on the type of responses. For continuous outcomes, the test statistic for randomization-based methods is difference in group means; and the t-test is used for population-based methods. For binary outcomes, the adopted test statistic is the linear-rank test statistic for randomization-based methods; and a χ^2-test is applied for population-based methods.

Table 4.1: *Homogeneous responses: missing data methods*

	Binary	Continuous
Randomization-based	the unconditional/conditional method, RBMI	the unconditional/conditional method. RBMI
Population-based	the "worst" method the "best-worst" method logistic regression imputation complete case analysis	the "worst" method the "best-worst" method mean imputation complete case analysis the maximum likelihood multiple imputation (stochastic)*

*: the stochastic regression imputation is used as the imputation technique in multiple imputation.

4.2 Methods Comparison when Responses are Continuous

4.2.1 Discussion about the Best-Worst Method and the Worst Methods

- The "best-worst" method

  After imputing missing outcomes by the "best-worst" method, a randomization test based on the linear rank test is adopted.

  - Severe type I error rate inflation occurred under all discussed procedures. This fact is presented in Figures 4.1. If a method's type I error rate is above the red dashed line, it indicates a type I error inflation case. The red dashed line is the upper bound of a type I error rate based on 1000 replications when the significance level is 0.05.

    To be more specific, if two treatments do not differ by assumption, applying the "best-worst" method will create an artificial dissimilarity in responses between groups since the missing in the beneficial (placebo) group is replaced by the

"worst" ("best") values, which explains the type I error rate inflation.

- When applying the "best-worst" method, type I error rates and power are sensitive to the overall missing proportion p_{ms}. When p_{ms} increases, type I error rate inflation become more severe since more missing values are imputed. The same logic can be applied to explain the significant drop in power when p_{ms} increases (this fact is presented in Figures 4.2).
- The "best-worst" method has the most severe type I error inflation problem among all methods discussed, and it is the most conservative method in terms of statistical power. Tables 4.2-4.7 present methods comparisons under different missing data mechanisms.
- In conclusion, when responses are continuous, the "best-worst" method is not recommended under all discussed procedures and missing data mechanisms due to its significant type I error inflation and conservative statistical power.

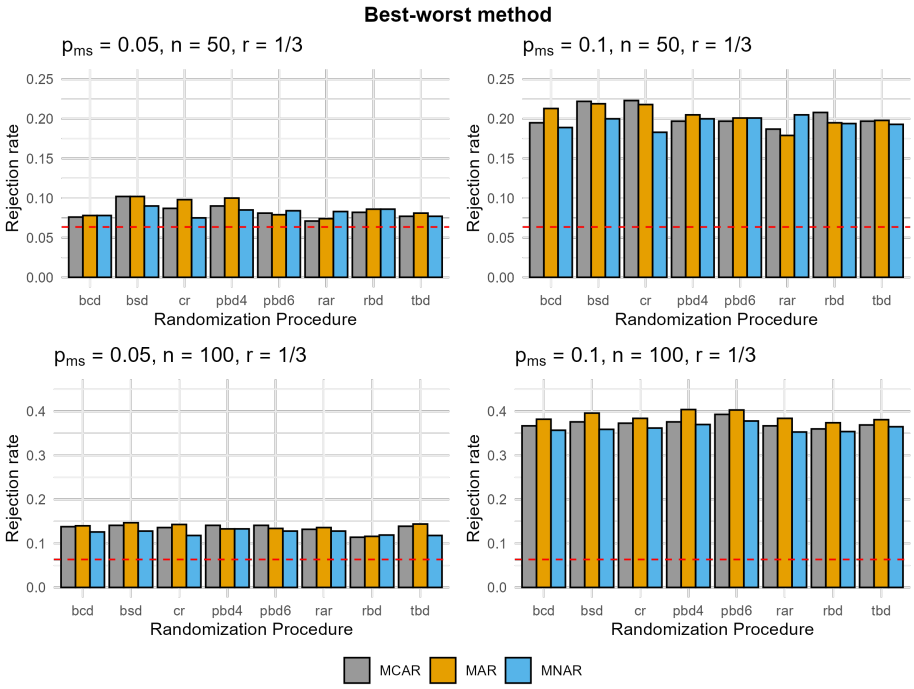

Figure 4.1: *Homogeneous responses (continuous): the best-worst method, type I error rates,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

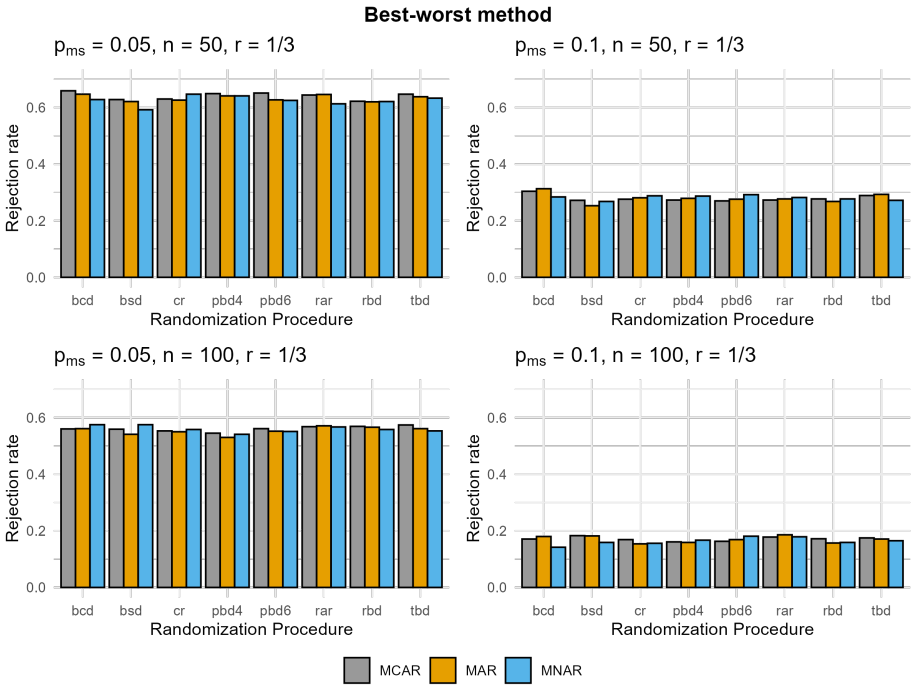

Figure 4.2: *Homogeneous responses (continuous): the best-worst method, power,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

- The "worst" method

  After imputing missing outcomes by the "worst" method, a randomization test based on the linear rank test is adopted.

  - Severe type I error rate inflation is observed under MAR, see Figure 4.3 for details. Based on the simulation settings, more missing outcomes occur in the beneficial group under MAR than the one under MCAR or MNAR. Missing values are replaced by values representing the worst treatment effect. Thus, if two treatments do not differ by assumption, when more missing values are imputed in one group than another one, an artificial dissimilarity between groups is created, which causes the type I error ratee inflation.
  - The statistical power is sensitive to the missing data mechanism and the missing proportion p_{ms}. More missing outcomes in the beneficial group are replaced by the "worst" values under MAR than MCAR and MNAR. Thus a smaller power is observed under MAR. The power drops more significantly when p_{ms} increases, see Figure 4.4 for details. The statistical power of the "worst" method ranks the second to the last (see Tables 4.2-4.7 for details).
  - In conclusion, when responses are continuous, the "worst" method is not recommended due to its inflated type I error rates and conservative power under MAR. Under MNAR, the "worst" method is not recommended because of its conservative power. As for MCAR, the "worst" method has smaller power than other methods.

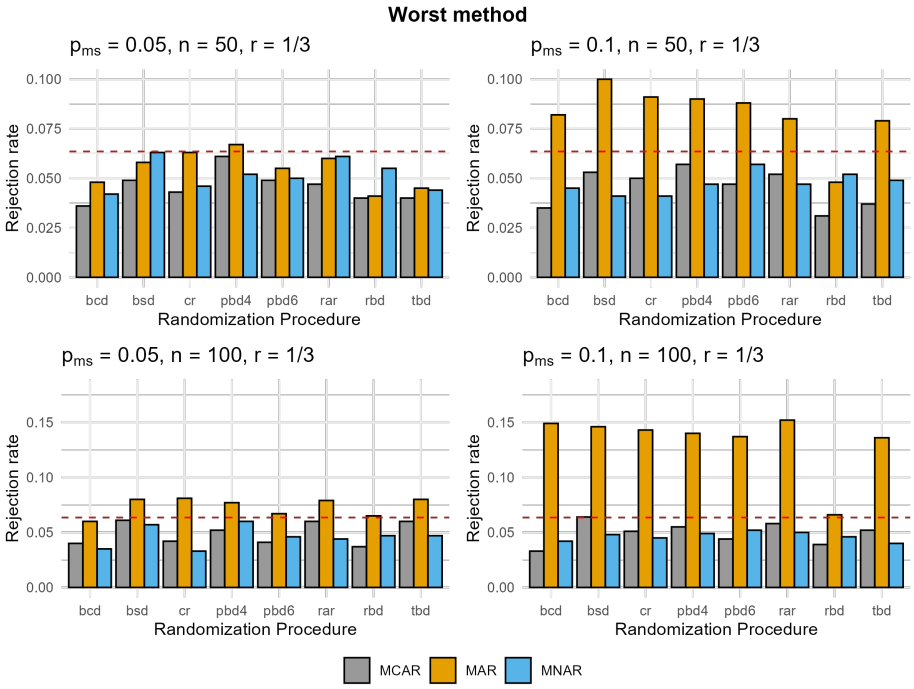

Figure 4.3: *Homogeneous responses (continuous): the worst method, type I error rates,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

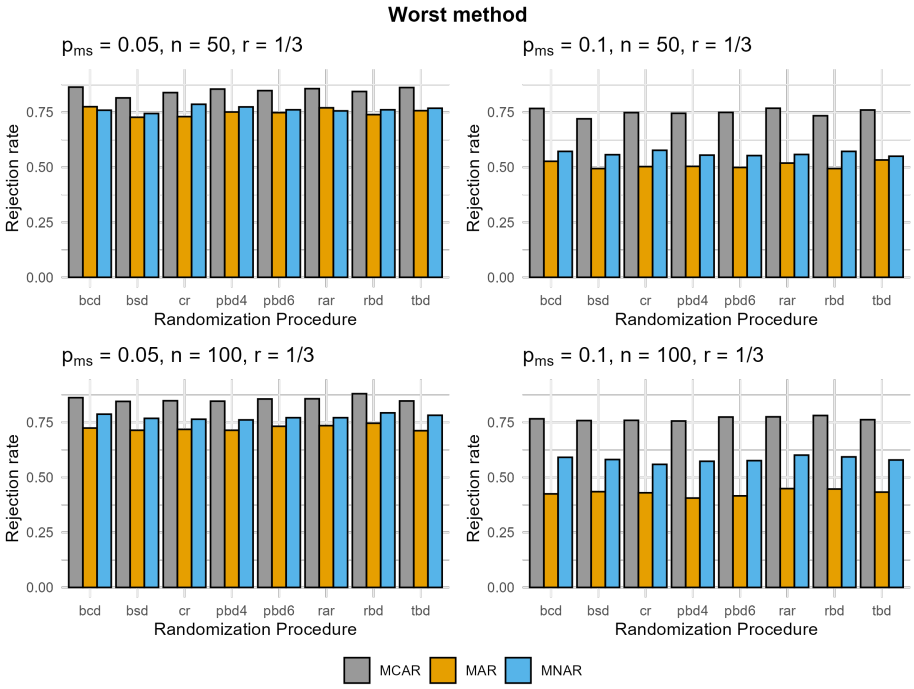

Figure 4.4: *Homogeneous responses (continuous): the worst method, power,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

4.2.2 Methods Comparison under MCAR

As discussed before, the "best-worst" method is are not recommended due to its poor performance in power and type I error rates. The "worst" method is not recommended because of its conservative power. As for the maximum likelihood method, it is not recommended because of its apparent type I error inflation problem under all discussed procedures. See Figure 4.5 for details.

Based on the discussion above, the "worst", the "best-worst" methods, and the maximum likelihood method are excluded from the comparison. The remaining methods have similar performance in power. Part of the simulation results are presented in Table 4.2 - 4.3 for illustration purposes. Results under other simulation cases are similar and they are available in the Appendix. From the type I error control perspective, the best procedure is the conditional reference set method. Under the BSD, all remaining techniques should be adopted with caution because of the potential type I error rate inflation problem. Similar type I error rate inflation under the BSD was also noted by Wang et al. (2020). More research is needed to explain this fact. Note that these inflated type I error rates are just outside the 95% upper bound for the type I rate for the number of replication in simulation.

Though the complete-case analysis is the quickest and easiest solution to missing data, it is only valid under the MCAR assumption. When a large number of missing values occur, significant power loss is expected under the complete-case analysis. Concerns about applying the complete-case analysis are discussed in Section 1.5.1.

Concerns about applying mean imputation are discussed by Van Buuren (2018). The mean imputation underestimates the variance, disturbs the relations between variables, and biases almost any estimates other than the mean when data are not MCAR.

In conclusion, the "best-worst", the "worst," and the maximum likelihood method are not recommended. All remaining methods are comparable in power for the designs studied. The conditional reference set method exhibits the best type I error rate control. The randomization-based methods can serve as alternatives for population-based methods in handling missing data based on their performance in power and type I error rate control.

Table 4.2: *Homogeneous responses (continuous): type I error rates (MCAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.367	0.033	0.046	0.050	0.046	0.058	0.046	0.048	0.047
2	BSD	0.376	0.064	0.062	0.063	0.062	0.077	0.061	0.048	0.061
3	CR	0.373	0.051	0.052	0.052	0.052	0.059	0.051	0.059	0.050
4	PBD(blocksize = 4)	0.376	0.055	0.057	0.061	0.058	0.074	0.059	0.050	0.056
5	PBD(blocksize = 6)	0.393	0.044	0.049	0.050	0.049	0.065	0.051	0.050	0.043
6	RAR	0.367	0.058	0.058	0.061	0.058	0.084	0.058	0.054	0.057
7	RBD	0.360	0.039	0.047	0.049	0.048	0.058	0.049	0.045	0.045
8	TBD	0.369	0.052	0.055	0.061	0.055	0.074	0.056	0.061	0.056

Table 4.3: *Homogeneous responses (continuous): power (MCAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.171	0.766	0.916	0.919	0.916	0.936	0.916	0.905	0.912
2	BSD	0.183	0.758	0.902	0.902	0.902	0.919	0.898	0.887	0.896
3	CR	0.169	0.759	0.896	0.901	0.898	0.926	0.901	0.909	0.895
4	PBD(blocksize = 4)	0.161	0.756	0.904	0.902	0.904	0.917	0.898	0.903	0.897
5	PBD(blocksize = 6)	0.163	0.774	0.907	0.910	0.907	0.927	0.902	0.892	0.902
6	RAR	0.178	0.775	0.900	0.907	0.902	0.927	0.903	0.899	0.901
7	RBD	0.172	0.781	0.911	0.911	0.911	0.930	0.910	0.910	0.905
8	TBD	0.175	0.762	0.901	0.904	0.901	0.915	0.899	0.897	0.878

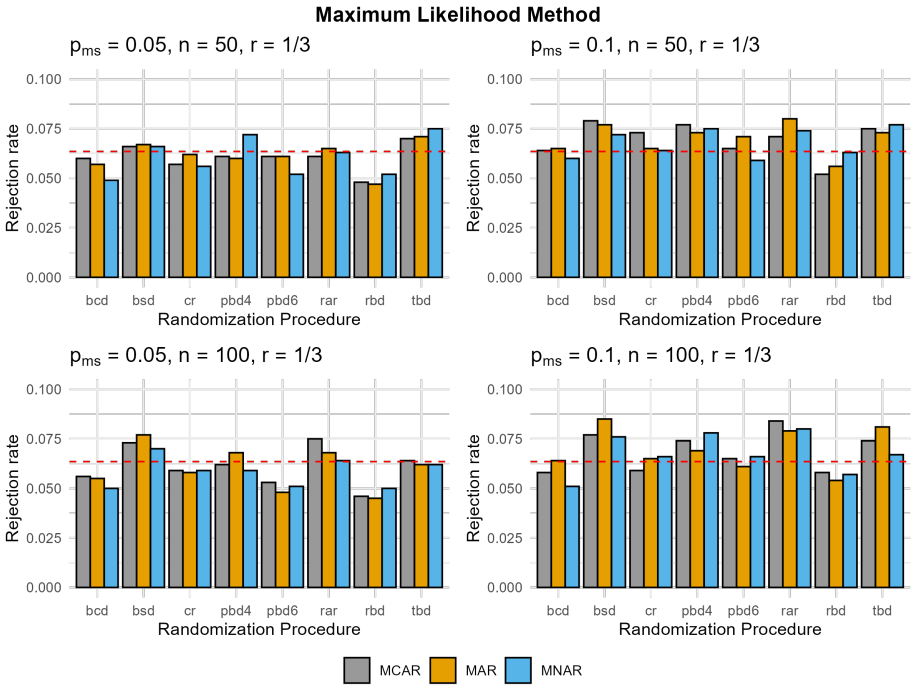

Figure 4.5: *Homogeneous responses (continuous): maximum likelihood, type I error rates,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

4.2.3 Methods Comparison under MAR

Under MAR, the "best-worst", the "worst" method, and the maximum likelihood method are not recommended because of their poor type I error rate control. The remaining methods' performance in power is similar, and slight inflation in type I error rates is observed under certain randomization procedures such as the BSD. The conditional reference set method is preferred since no type I error rate inflation is observed for procedures studied. Details for the comparison of methods are available in Tabld 4.4 - 4.5. More results are available in the Appendix. Note that applying mean imputation or the complete-case analysis

when the data are not MCAR potentially risks the statistical inference from multiple perspectives, e.g., underestimating or overestimating the estimator's variance, etc. The section 1.5.1, 1.5.2 and Van Buuren (2018) discussed potential problems of applying mean imputation (or the complete-case analysis) in detail. Multiple imputation is a valid parametric method to handle missing data under MAR.

Choosing an appropriate missing data method depends on the randomization procedure used in the trial since some slight type I error rate inflation is observed under a particular combination of missing data method and randomization procedure. Based on simulation results, the unconditional/unconditional reference set, multiple imputation, and RBMI are recommended under MAR. The complete-case analysis and the mean imputation are less recommended because of potential problems. The "best-worst", the "worst" method, and the maximum likelihood are not recommended.

Table 4.4: *Homogeneous responses (continuous): type I error rates (MAR,* $n = 100$*,* $p_{ms} = 0.1$*)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.382	0.149	0.049	0.053	0.048	0.064	0.048	0.055	0.051
2	BSD	0.396	0.146	0.064	0.069	0.063	0.085	0.066	0.056	0.061
3	CR	0.384	0.143	0.052	0.050	0.052	0.065	0.052	0.052	0.050
4	PBD(blocksize = 4)	0.404	0.140	0.055	0.055	0.055	0.069	0.054	0.047	0.053
5	PBD(blocksize = 6)	0.403	0.137	0.046	0.047	0.046	0.061	0.044	0.051	0.041
6	RAR	0.384	0.152	0.063	0.067	0.063	0.079	0.064	0.059	0.063
7	RBD	0.374	0.066	0.043	0.042	0.043	0.054	0.042	0.041	0.039
8	TBD	0.381	0.136	0.061	0.062	0.060	0.081	0.062	0.058	0.054

Table 4.5: *Homogeneous responses (continuous): power (MAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.180	0.425	0.911	0.905	0.910	0.929	0.908	0.905	0.907
2	BSD	0.182	0.435	0.900	0.895	0.899	0.905	0.893	0.877	0.891
3	CR	0.154	0.430	0.903	0.901	0.902	0.925	0.902	0.902	0.895
4	PBD(blocksize = 4)	0.159	0.406	0.907	0.907	0.906	0.925	0.906	0.904	0.904
5	PBD(blocksize = 6)	0.169	0.416	0.917	0.914	0.917	0.930	0.910	0.898	0.906
6	RAR	0.186	0.449	0.903	0.907	0.903	0.920	0.903	0.905	0.900
7	RBD	0.157	0.447	0.914	0.909	0.914	0.932	0.911	0.919	0.911
8	TBD	0.171	0.433	0.896	0.893	0.896	0.916	0.891	0.890	0.863

4.2.4 Methods Comparison under MNAR

As discussed in Section 1.5.5, none of the methods we discussed so far can deal with MNAR. Particular assumptions about the missing mechanism are required to proceed with data analyses under MNAR. To ascertain the exact missing data mechanism require follow-up data from non-respondents, which is impractical in most cases. Thus, all missing mechanisms are somehow non-testable in practice. However, we can compare the different methods' performance under MNAR from simulation studies.

If we have strong confidence in MNAR, the data analysis should be adapted based on the reason behind the missing values. If not confident with the exact missing mechanism, the methods discussed are still available tools to handle missing data. The simulation results can be considered as a reference before choosing a missing data method under MNAR. Due to type I error rate inflation, the "best-worst", the "worst" methods, and maximum likelihood are not recommended. The remaining methods have similar performance in power. As for the type I error rate, most methods have minor inflation problems under some procedures (e.g., BSD). Generally speaking, randomization-based methods are comparable to parametric methods in type I error rates and power. Part of the simulation results are

available in Tables 4.6. - 4.7. Results under other simulation cases are similar and are available in the Appendix. Randomization-based methods are comparable to population-based methods in handling missingness.

Table 4.6: *Homogeneous responses (continuous): type I error rates (MNAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.357	0.042	0.037	0.037	0.037	0.051	0.038	0.035	0.035
2	BSD	0.359	0.048	0.060	0.059	0.060	0.076	0.064	0.055	0.061
3	CR	0.362	0.045	0.049	0.049	0.049	0.066	0.047	0.055	0.046
4	PBD(blocksize = 4)	0.370	0.049	0.058	0.059	0.058	0.078	0.063	0.051	0.059
5	PBD(blocksize = 6)	0.378	0.052	0.049	0.049	0.049	0.066	0.046	0.059	0.045
6	RAR	0.353	0.050	0.063	0.066	0.063	0.080	0.064	0.051	0.064
7	RBD	0.354	0.046	0.042	0.046	0.043	0.057	0.043	0.043	0.044
8	TBD	0.365	0.040	0.057	0.056	0.057	0.067	0.055	0.062	0.056

Table 4.7: *Homogeneous responses (continuous): power (MNAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.142	0.591	0.893	0.891	0.893	0.910	0.891	0.886	0.891
2	BSD	0.159	0.581	0.876	0.876	0.877	0.899	0.880	0.873	0.875
3	CR	0.156	0.559	0.872	0.876	0.872	0.897	0.873	0.875	0.870
4	PBD(blocksize = 4)	0.167	0.573	0.882	0.878	0.882	0.905	0.876	0.899	0.877
5	PBD(blocksize = 6)	0.181	0.576	0.890	0.891	0.890	0.916	0.886	0.884	0.883
6	RAR	0.179	0.601	0.888	0.888	0.887	0.911	0.883	0.887	0.875
7	RBD	0.159	0.593	0.894	0.892	0.894	0.920	0.889	0.898	0.885
8	TBD	0.165	0.579	0.886	0.886	0.886	0.908	0.886	0.884	0.848

4.3 Methods Comparison when Responses are Binary

4.3.1 Discussion about the Best-Worst and the Worst Methods

After imputing missing responses by the "best-worst" method, a randomization test based on the linear-rank test is adopted. Note that for missing responses in a beneficial group, i.e., the treatment A in our simulation settings, imputed values for missing responses are failures. If missing responses are in the less beneficial group, i.e., the treatment B, imputed values are successes.

- The "best-worst" method
  - Severe type I error rate inflation occurred under all discussed procedures. The fact is presented in Figure 4.6. Similar type I error rate inflation is observed for continuous responses. In contrast, no such severe type I error rate inflation is observed under other methods. This fact is presented in Tables 4.8, 4.10 and 4.12. Applying the "best-worst" method will create an artificial dissimilarity in responses between groups since missing values in the beneficial (placebo) group are replaced by failures (success) if two treatments do not differ by assumption. It causes type I error rate inflation.
  - When applying the "best-worst" method, type I error rates and power are sensitive to the overall missing proportion p_{ms}. When p_{ms} increases, the type I error rate inflation becomes more severe since more missing values are imputed. The same logic can be applied to explain the significant drop in power when p_{ms} increases (this fact is presented in Figure 4.7).
  - The "best-worst" method has the most severe type I error rate inflation problem among all methods discussed, and it is the most conservative method in terms of statistical power. Tables 4.8 - 4.13 present methods comparisons under different missing data mechanisms.

– In conclusion, when responses are binary, the "best-worst" method is not recommended under all discussed procedures and missing data mechanisms due to its significant type I error rate inflation and conservative statistical power.

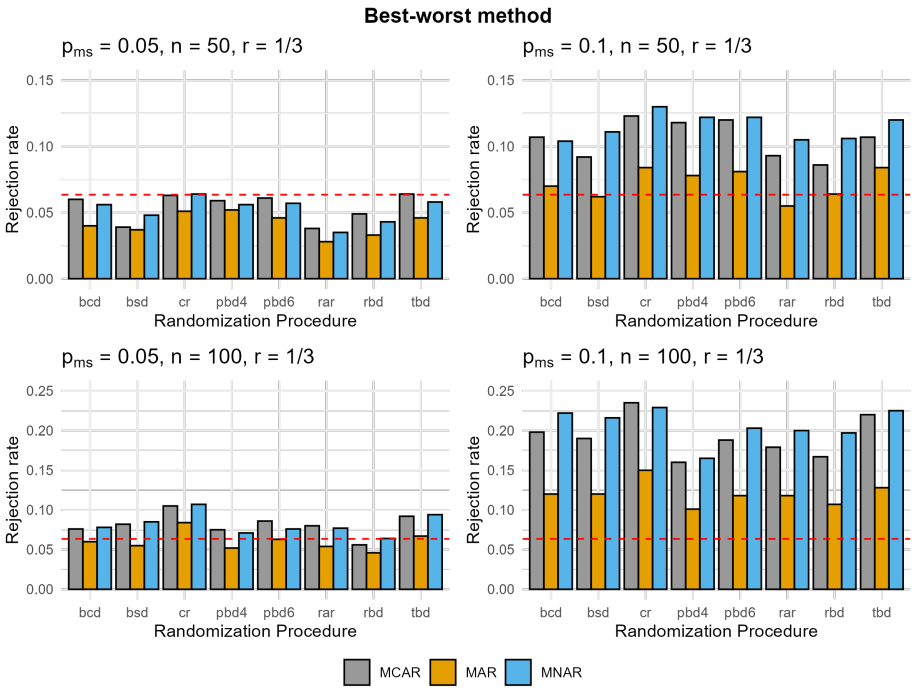

Figure 4.6: *Homogeneous responses (binary): the best-worst method, type I error rates,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

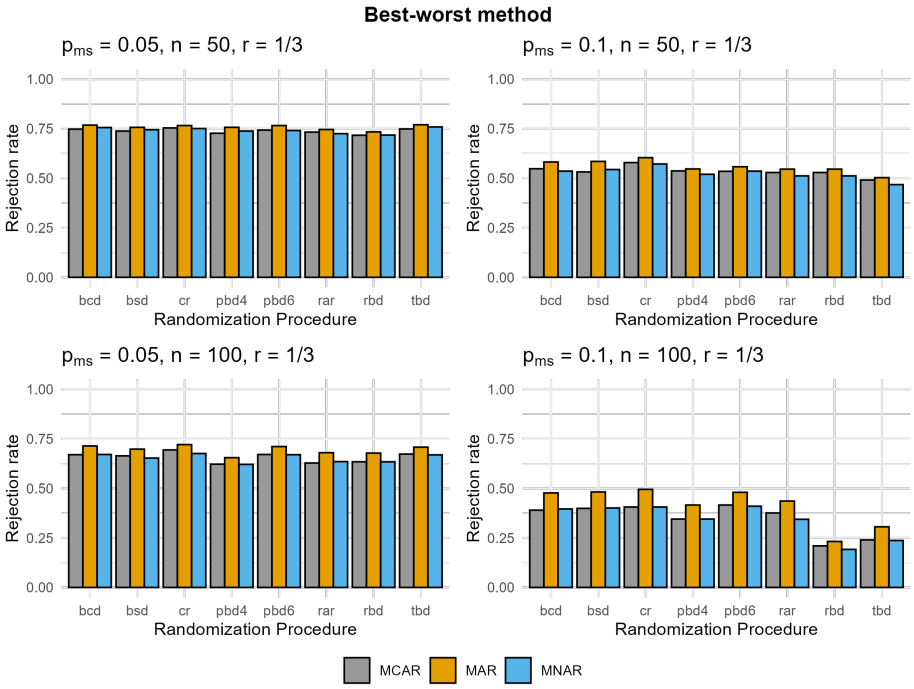

Figure 4.7: *Homogeneous responses (binary): the best-worst method, power,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

- The "worst" method

  After imputing missing outcomes by the "worst" method, a randomization test based on the linear-rank test statistic is adopted. Imputed missing outcomes are failures in both two treatment groups.

  - Type I error rate inflation is only spotted in few cases. However, the statistical power is sensitive to the missing data mechanism and the missing proportion p_{ms}. This fact is presented in Figures 4.8-4.9. More missing responses in the beneficial group are replaced by the "worst" values under MAR than MCAR and MNAR. Thus, a smaller power is observed under MAR. Also, the power drops more significantly when p_{ms} increases; the statistical power of the "worst" method ranks the second to the last(see Tables 4.9, 4.11 and 4.13 for details).
  - In conclusion, the "worst" method is not recommended due to its relatively small power under MAR and MNAR for binary responses. As for MCAR, the "worst" method has smaller power than other methods. Also, it should be adopted with caution when the missing proportion p_{ms} is large.

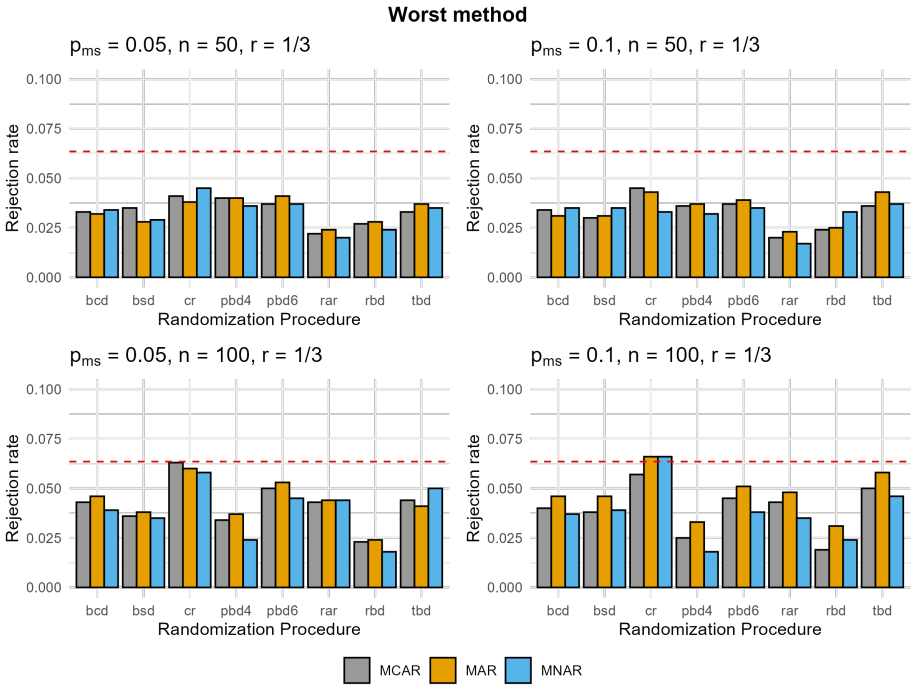

Figure 4.8: *Homogeneous responses (binary): the worst method, type I error rates,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

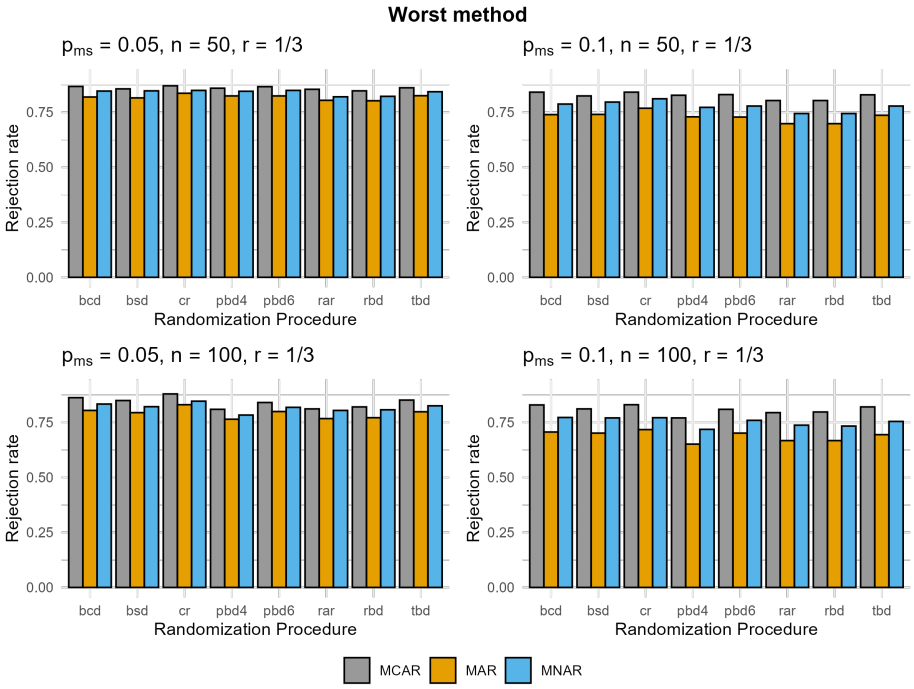

Figure 4.9: *Homogeneous responses (binary): the worst method, power,* $p_{ms} = 0.05$ *(left plots),* $p_{ms} = 0.1$ *(right plots).*

4.3.2 Methods Comparison under MCAR

Under MCAR, the "best-worst" method is not recommended because of its severe type I error rate inflation and conservative power. Thus, the "best-worst" method is excluded from the discussion. Part of the simulation results of power and type I error rates are presented in Tables 4.8 and 4.9 for illustration purposes. More results are available in the Appendix and the pattern is similar. Generally, considerations from multiple aspects are needed when selecting a method, such as the adopted randomization procedure, the missing proportion, the focus on type I error rate control or the statistical power, etc. Simulation results above show that proposed randomization-based methods are comparable to parametric methods in terms of power and type I error rates under the MCAR assumption.

Table 4.8: *Homogeneous responses (binary): type I error rates (MCAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.198	0.040	0.030	0.047	0.051	0.050	0.065
2	BSD	0.190	0.038	0.032	0.050	0.041	0.041	0.039
3	CR	0.235	0.057	0.041	0.086	0.060	0.059	0.058
4	PBD(blocksize = 4)	0.160	0.025	0.040	0.052	0.050	0.038	0.030
5	PBD(blocksize = 6)	0.188	0.045	0.039	0.058	0.050	0.032	0.047
6	RAR	0.179	0.043	0.046	0.060	0.054	0.043	0.047
7	RBD	0.167	0.019	0.018	0.043	0.032	0.029	0.023
8	TBD	0.220	0.050	0.041	0.058	0.049	0.040	0.045

Table 4.9: *Homogeneous responses (binary): power (MCAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.390	0.829	0.817	0.831	0.859	0.857	0.889
2	BSD	0.399	0.811	0.813	0.842	0.832	0.834	0.830
3	CR	0.406	0.830	0.835	0.858	0.856	0.858	0.856
4	PBD(blocksize = 4)	0.345	0.770	0.824	0.832	0.845	0.828	0.803
5	PBD(blocksize = 6)	0.416	0.809	0.833	0.841	0.849	0.846	0.835
6	RAR	0.376	0.794	0.827	0.823	0.845	0.853	0.832
7	RBD	0.210	0.797	0.811	0.839	0.842	0.831	0.820
8	TBD	0.240	0.820	0.840	0.856	0.852	0.841	0.843

4.3.3 Methods Comparison under MAR

The "best-worst" and the "worst" methods are not recommended due to their conservative power and inflated type I error rates (under the "best-worst" method only). Thus these two methods are excluded from the discussion below.

Part of simulation results of power and type I error rates are presented in Table 4.10 and 4.11 for illustration purposes. More results are available in Appendix and the pattern is similar. The complete-case analysis is recommended. However, the complete-case analysis is only valid under MCAR, and it has a potential power loss trend when the overall missing proportion p_{ms} increases. Potential problems of applying the complete-case analysis are discussed before. In practice, the complete-case analysis should be adopted with caution under MAR since the situation might be more complex than the simulated scenario. When a procedure other than CR or BCD is adopted, randomization-based and parametric methods have similar performance in type I error rates and power.

Table 4.10: *Homogeneous responses (binary): type I error rates (MAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.120	0.046	0.031	0.055	0.052	0.047	0.067
2	BSD	0.120	0.046	0.028	0.054	0.042	0.041	0.037
3	CR	0.150	0.066	0.045	0.076	0.061	0.065	0.063
4	PBD(blocksize = 4)	0.101	0.033	0.038	0.048	0.049	0.042	0.030
5	PBD(blocksize = 6)	0.118	0.051	0.044	0.057	0.056	0.036	0.051
6	RAR	0.118	0.048	0.044	0.053	0.057	0.045	0.043
7	RBD	0.107	0.031	0.014	0.029	0.035	0.028	0.021
8	TBD	0.128	0.058	0.040	0.061	0.050	0.044	0.045

Table 4.11: *Homogeneous responses (binary): power (MAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.477	0.706	0.812	0.844	0.871	0.868	0.893
2	BSD	0.482	0.701	0.828	0.856	0.848	0.847	0.844
3	CR	0.495	0.717	0.841	0.855	0.862	0.864	0.867
4	PBD(blocksize = 4)	0.416	0.651	0.821	0.830	0.841	0.824	0.805
5	PBD(blocksize = 6)	0.480	0.701	0.825	0.844	0.845	0.848	0.827
6	RAR	0.436	0.667	0.824	0.842	0.848	0.850	0.832
7	RBD	0.232	0.667	0.800	0.838	0.839	0.827	0.815
8	TBD	0.306	0.694	0.849	0.860	0.863	0.846	0.850

4.3.4 Methods Comparison under MNAR

The "best-worst" method is not recommended due to its conservative power and severe type I error rate inflation. Thus, the "best-worst" method is excluded from the discussion

below. The "worst" method is less recommended than the remaining methods since it has relatively small statistical power. The complete-case analysis is recommended in our simulation scenario. However, in practice, the complete-case analysis should be adopted with cautions; reasons are discussed previously.

In conclusion, randomization-based methods are comparable to parametric methods under MNAR when responses are binary.

Table 4.12: *Homogeneous responses (binary): type I error rates (MNAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.222	0.037	0.022	0.050	0.048	0.043	0.059
2	BSD	0.216	0.039	0.035	0.048	0.047	0.045	0.043
3	CR	0.229	0.066	0.045	0.077	0.063	0.063	0.063
4	PBD(blocksize = 4)	0.165	0.018	0.027	0.043	0.044	0.025	0.019
5	PBD(blocksize = 6)	0.203	0.038	0.043	0.057	0.049	0.030	0.048
6	RAR	0.200	0.035	0.043	0.063	0.053	0.039	0.043
7	RBD	0.197	0.024	0.018	0.046	0.038	0.032	0.030
8	TBD	0.225	0.046	0.035	0.070	0.042	0.037	0.037

Table 4.13: *Homogeneous responses (binary): power (MNAR, $n = 100$, $p_{ms} = 0.1$)*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.396	0.772	0.788	0.821	0.841	0.838	0.876
2	BSD	0.401	0.770	0.801	0.818	0.824	0.826	0.823
3	CR	0.406	0.771	0.810	0.837	0.842	0.844	0.842
4	PBD(blocksize = 4)	0.345	0.718	0.804	0.819	0.826	0.801	0.785
5	PBD(blocksize = 6)	0.410	0.759	0.806	0.829	0.823	0.811	0.809
6	RAR	0.344	0.737	0.807	0.812	0.825	0.832	0.816
7	RBD	0.192	0.733	0.7777	0.803	0.821	0.802	0.797
8	TBD	0.237	0.754	0.823	0.852	0.838	0.822	0.826

Chapter 5: Results Under Heterogeneity

In this chapter, the performance of randomization-based methods is evaluated by comparing them with population-based methods from these three following perspectives:

- patients' responses are affected by time trends,
- outliers exist in patients' responses,
- missing proportions increase over time.

Due to the poor performance of the "best-worst" method and the "worst" method, they are excluded from the following discussion.

5.1 Time Trends in Responses

5.1.1 Missing Data Methods

The methods studied are listed in the Table 5.1.

5.1.2 Simulation Results

When the responses are continuous, the complete-case analysis, mean imputation, and multiple imputation are tempting to use in practice due to their simplicity and availability. However, these three methods essentially rely on a two-sample t-test (in this thesis's simulations); the performance of these methods under a time trend deserves further investigation. A time trend is often observed in the clinical trial due to the sequential recruitment of patients, and it is not perceived before the start of the trial. The existence of a time trend may bias statistical inference when comparing two treatment effects. The performance of the methods is investigated when time trends exist in the following.

Table 5.1: *Time trend: missing data methods*

	Binary	Continuous
Randomization-based	the unconditional/conditional method, RBMI	the unconditional/conditional method RBMI
Population-based	logistic regression imputation complete case analysis	mean imputation complete case analysis maximum likelihood multiple imputation (stochastic)

Table 5.2: *Time trend: parameters in responses*

	Type I error rate	Power (n=50)	Power (n=100)
jth patient's responses	A: N(0.2+$\Delta_\theta(j)$,1) B: N(0.2+$\Delta_\theta(j)$,1)	A: N(1.2+$\Delta_\theta(j)$,1) B: N(0.2+$\Delta_\theta(j)$,1)	A: N(0.9+$\Delta_\theta(j)$,1) B: N(0.2+$\Delta_\theta(j)$,1)

The influence of a time trend is reflected in the responses' theoretical mean. First, let us define the amount of mean shift on jth patient's response $\Delta_\theta(j)$, where $\Delta_\theta(j) = (j-1)\theta/n$; where j is the position of jth patient in the randomization sequence, n is the total number of patients, and θ is the theoretical maximum mean shift on patient's response. Possible values of θ are (-2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2). Once θ is chosen, the mean shift on each patient's responses can be calculated. The later the patient joins the trial, the larger the influence (mean shift) on his/her response due to the time trend. The normal distribution is used to simulate patients's responses, see Table 5.2 for details.

Simulation results in the following sections are type I error rates and power based on 1000 replications when evaluating treatment effects. The way to simulate responses with missing

values is consistent with the previous case (available in Section 3.1) when no heterogeneity existed in patients' responses.

5.1.2.1 Type I Error Rates

Figures 5.1 - 5.8 and Table 5.3 below describe the type I error rate inflation among different methods and procedures. For each method under a fixed missing mechanism and randomization procedure (e.g., CR under MCAR when multiple imputation is adopted at $\theta = 2$), multiple type I error rates are simulated via different combination of simulation parameters such as the missing proportion $p_{ms} = 0.05$ or 0.1 and number of patients $n = 50$, 100. Part of the simulation results ($n = 100$, $p_{ms} = 0.1$) are presented in this section. More figures under other simulation cases are available in the Appendix. The smooth curve within each subplot is to indicate the trend of type I error rates or power. Beyond the red dashed line represents type I error rate inflation beyond simulation error.

In Table 5.3, the count in each cell is the total frequency of type I error rate inflation out of 8 simulation cases generated from different parameter Δs (i.e., different amount of time trends) under the case when the sample size $n = 100$ and the overall missing proportion $p_{ms} = 0.1$.

The following is a summary of simulation results for type I error rates under time trends in responses.

- Type I error rate inflation occurs when applying the maximum likelihood method among most randomization procedures.

- TBD is a randomization procedure that may have a sequence where one treatment is dominant in the second half; thus, it tends to be highly affected by the time trend. Figure 5.1 presents that parametric methods are poorer in type I error rates control than randomization-based methods under TBD regardless of the missing mechanism. Moreover, when the influence of time trend becomes larger (measured by the $|\theta|$), the inflation becomes more severe under parametric methods.

- In terms of the number of inflated cases, the randomization-based methods present more strict type I error rate control than the other methods discussed in simulations; see Table 5.3 for details.

- Block-based designs (such as RBD and PBD) exhibit better performance in type I error rate control than designs are not block-based when responses are affected by time trend; see Table 5.3 for details.

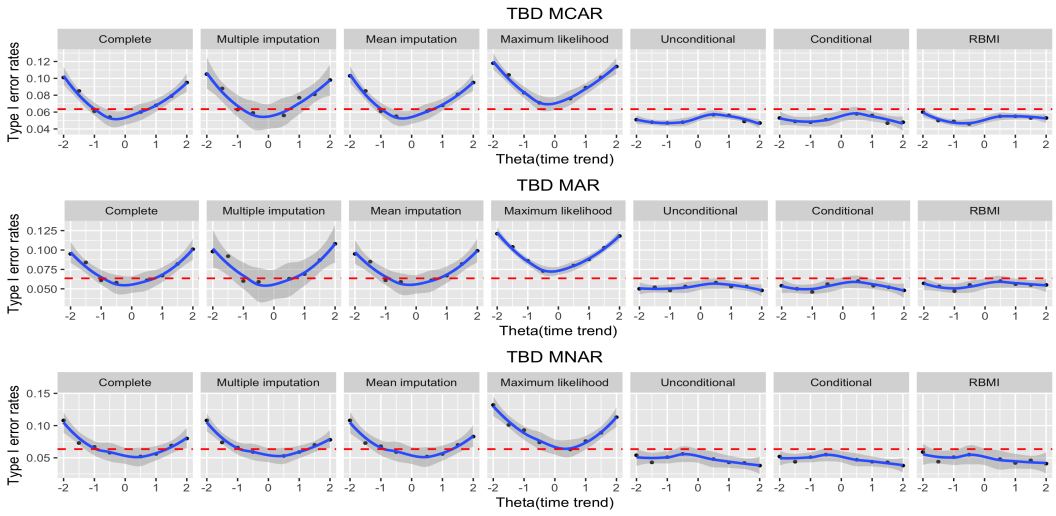

Figure 5.1: *Time trend: TBD, type I error rates ($p_{ms} = 0.1$, $n = 100$)*

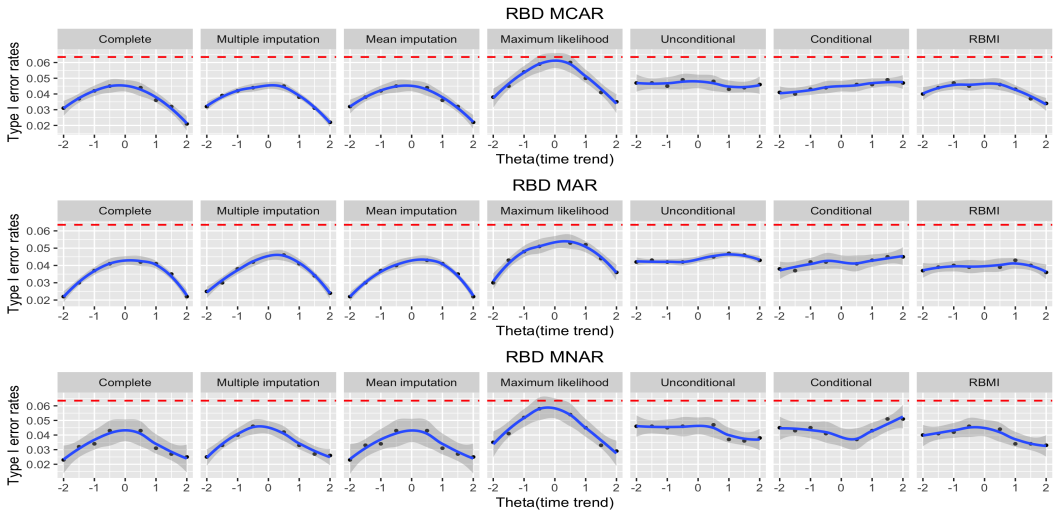

Figure 5.2: *Time trend: RBD (maximum blocksize = 6), type I error rates ($p_{ms} = 0.1$, $n = 100$)*

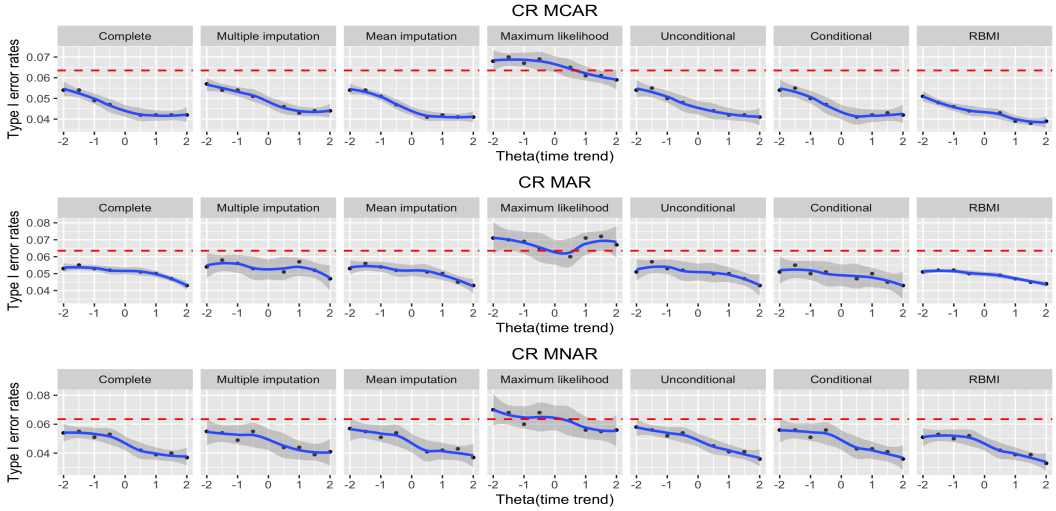

Figure 5.3: *Time trend: CR, type I error rates ($p_{ms} = 0.1$, $n = 100$)*

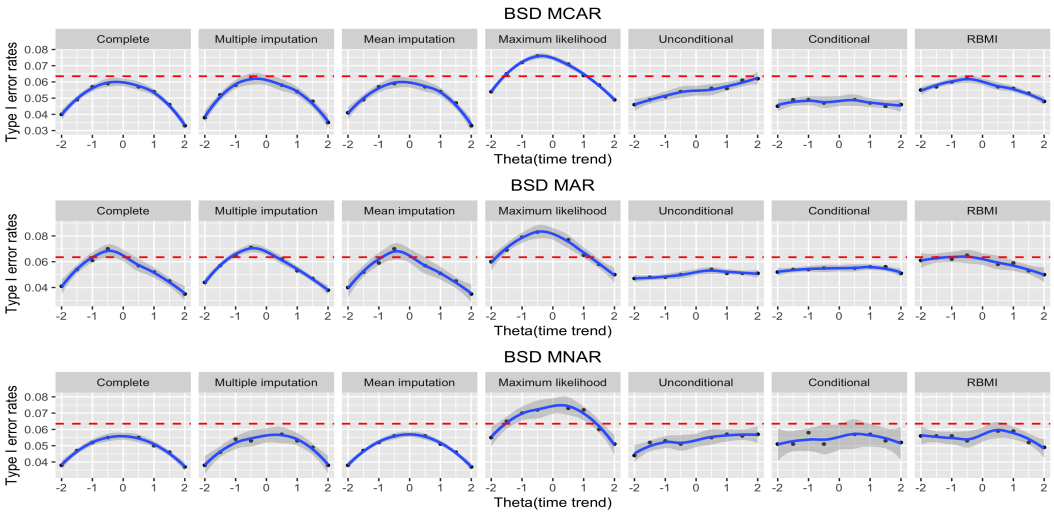

Figure 5.4: *Time trend: BSD, type I error rates ($p_{ms} = 0.1$, $n = 100$)*

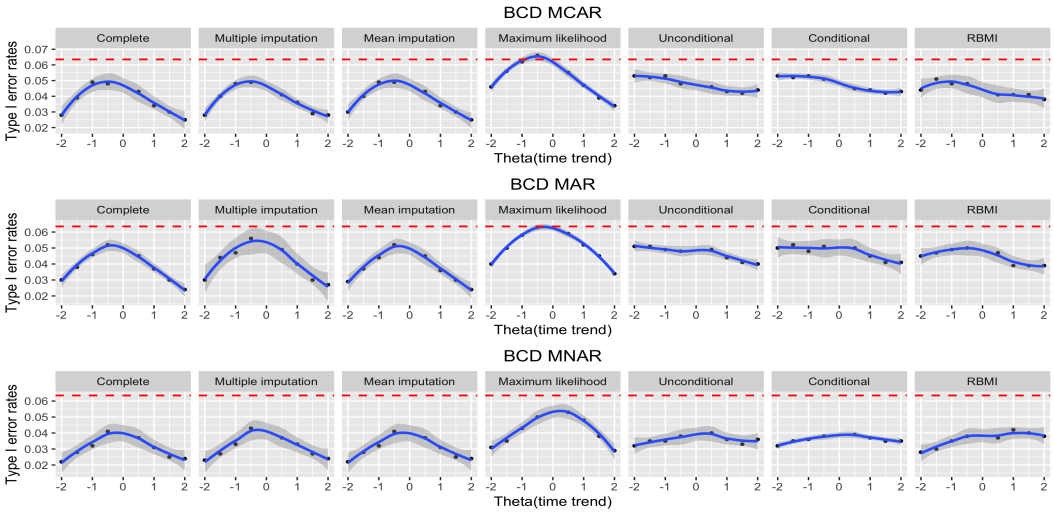

Figure 5.5: *Time trend: BCD, type I error rates ($p_{ms} = 0.1$, $n = 100$)*

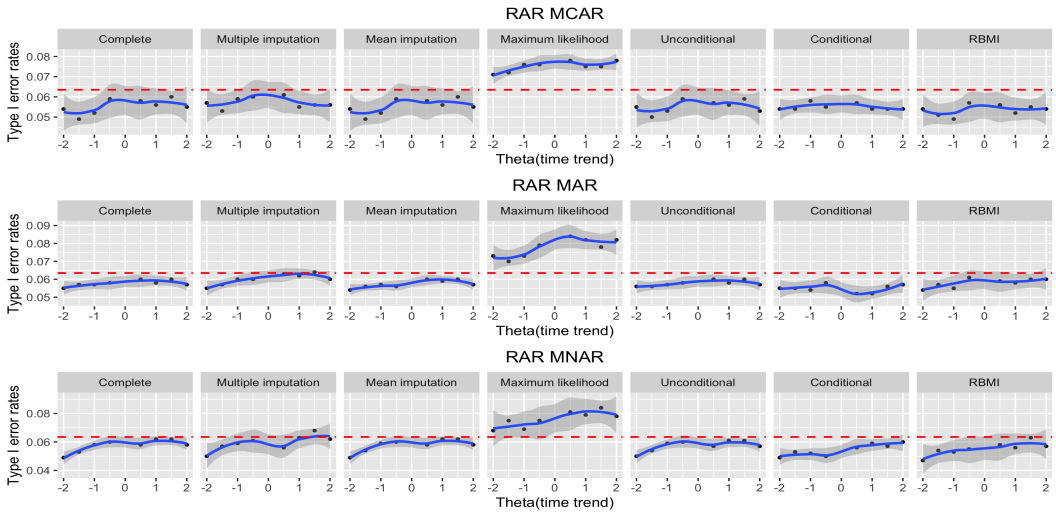

Figure 5.6: *Time trend: RAR, type I error rates ($p_{ms} = 0.1$, $n = 100$)*

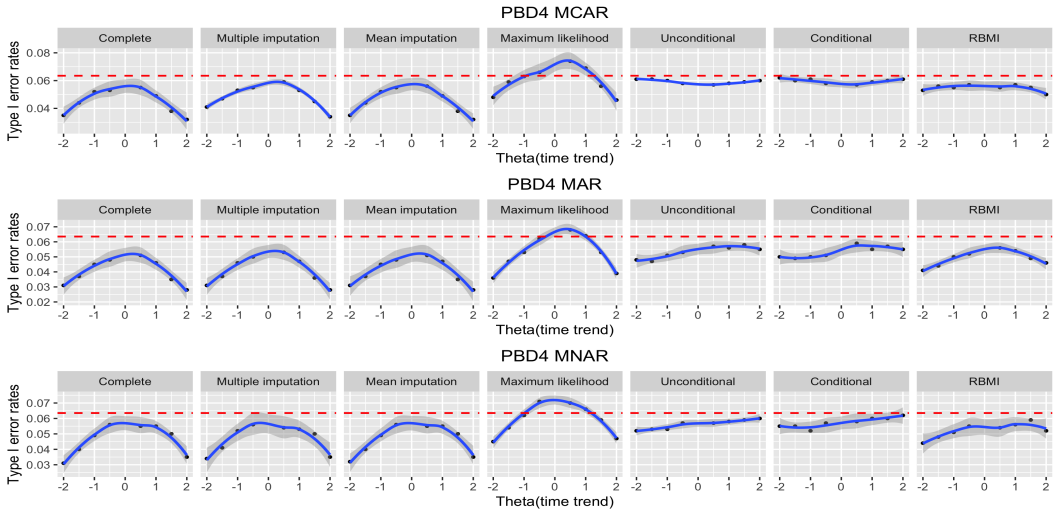

Figure 5.7: *Time trend: PBD (blocksize = 4), type I error rates ($p_{ms} = 0.1$, $n = 100$)*

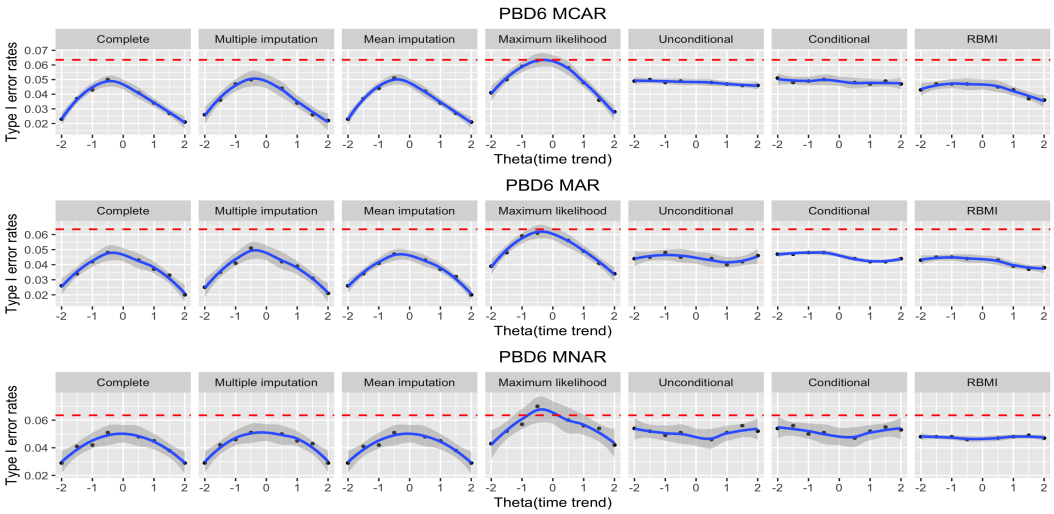

Figure 5.8: *Time trend: PBD (blocksize = 6), type I error rates ($p_{ms} = 0.1$, $n = 100$)*

Table 5.3: *Time Trends: Number of replications in which type I error rate is inflated out of eight simulation cases (no block designs vs. block designs; below the dashed line are block designs ($n = 100$, $p_{ms} = 0.1$)*

Missing Mechanisms	Randomization procedures	Conditional	Unconditional	Complete-case analysis	Multiple Imputation (stochastic)	The mean imputation	Maximum likelihood	RBMI
	BCD	0	0	0	0	0	1	0
	CR	0	0	0	0	0	5	0
MCAR	BSD	0	0	0	0	0	5	0
	RAR	0	0	0	0	0	8	0
	TBD	0	0	5	5	5	8	0
	BCD	0	0	0	0	0	0	0
	CR	0	0	0	0	0	7	0
MAR	BSD	0	0	1	2	1	5	1
	RAR	0	0	0	1	0	8	0
	TBD	0	0	5	5	5	8	0
	BCD	0	0	0	0	0	0	0
	CR	0	0	0	0	0	3	0
MNAR	BSD	0	0	0	0	0	5	0
	RAR	0	0	0	1	0	8	0
	TBD	0	0	5	5	5	7	0
	RBD	0	0	0	0	0	0	0
MCAR	PBD4	0	0	0	0	0	3	0
	PBD6	0	0	0	0	0	0	0
	RBD	0	0	0	0	0	0	0
MAR	PBD4	0	0	0	0	0	2	0
	PBD6	0	0	0	0	0	0	0
	RBD	0	0	0	0	0	0	0
MNAR	PBD4	0	0	0	0	0	3	0
	PBD6	0	0	0	0	0	1	0

5.1.2.2 Different Power Trends

Generally the maximum likelihood approach has the highest power among all methods. Under the randomization procedures RBD, BSD, BCD and PBD, we see an apparent trend in power; the trend is seen in Figure 5.9 for illustration purposes. More similar results are available in the Appendix. Population-based methods present a more apparent power loss compared to randomization-based methods when the influence of time trends becomes larger (measured by $|\theta|$).

When the randomization procedures are CR, RAR, and TBD, the maximum likelihood method maintains the highest power; methods relying on the t-test are less powerful. However, it is not observed that randomization-based methods have a smaller power loss than other population-based methods when the time trend's influence increases. The result under CR is presented in Figure 5.14. Similar trends are available in Figures 5.14-5.16. Consistent trends are observed in Rosenberger and Lachin (2016) where they compared the power of the randomization test and the t-test under different randomization procedures when responses are affected by linear time trends.

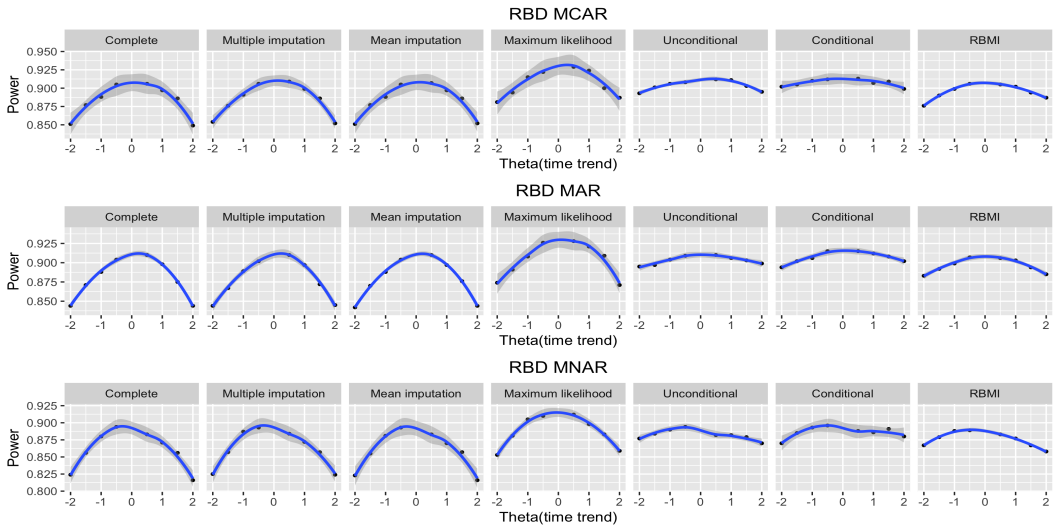

Figure 5.9: *Time trend: RBD, power ($p_{ms} = 0.1$, $n = 100$)*

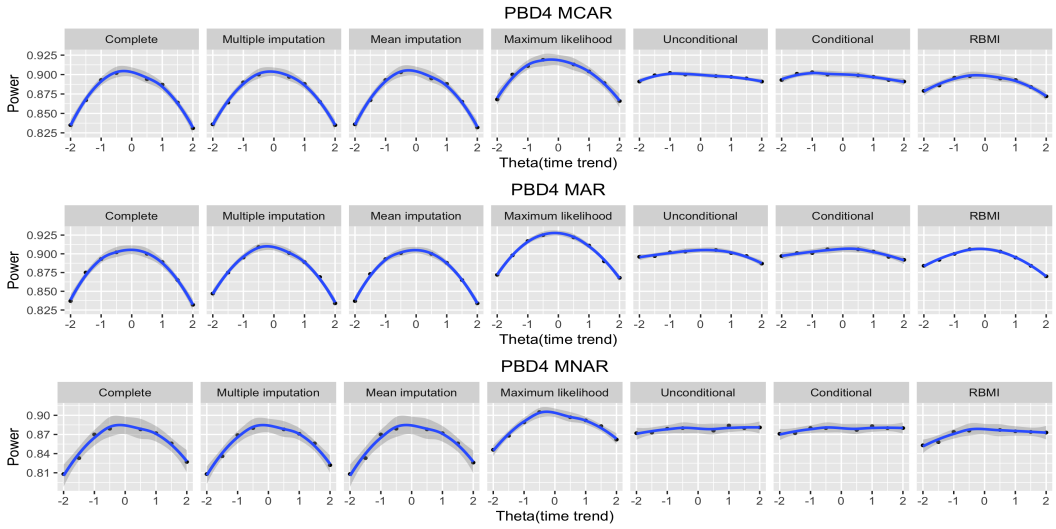

Figure 5.10: *Time trend: PBD (blocksize = 4), power ($p_{ms} = 0.1$, $n = 100$)*

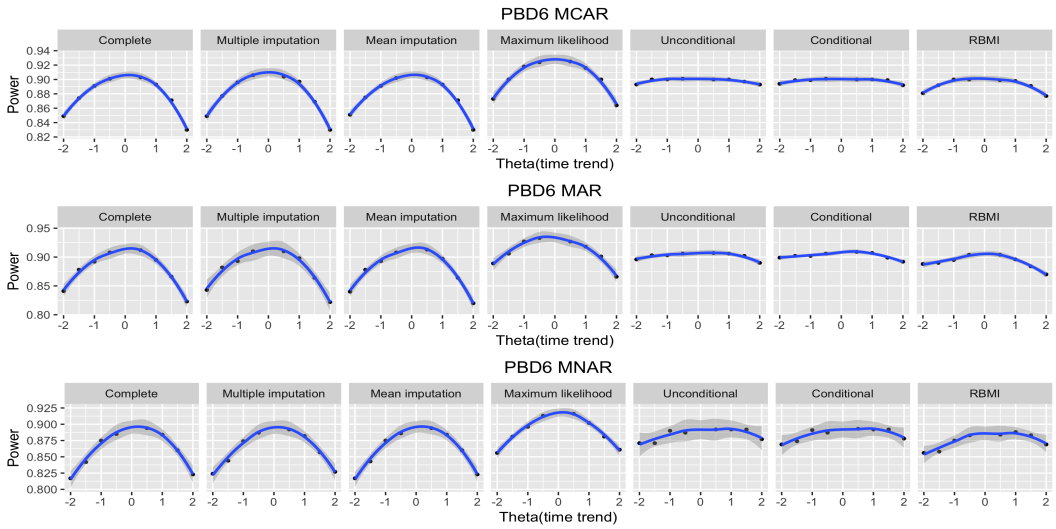

Figure 5.11: *Time trend: PBD (blocksize = 6), power ($p_{ms} = 0.1$, $n = 100$)*

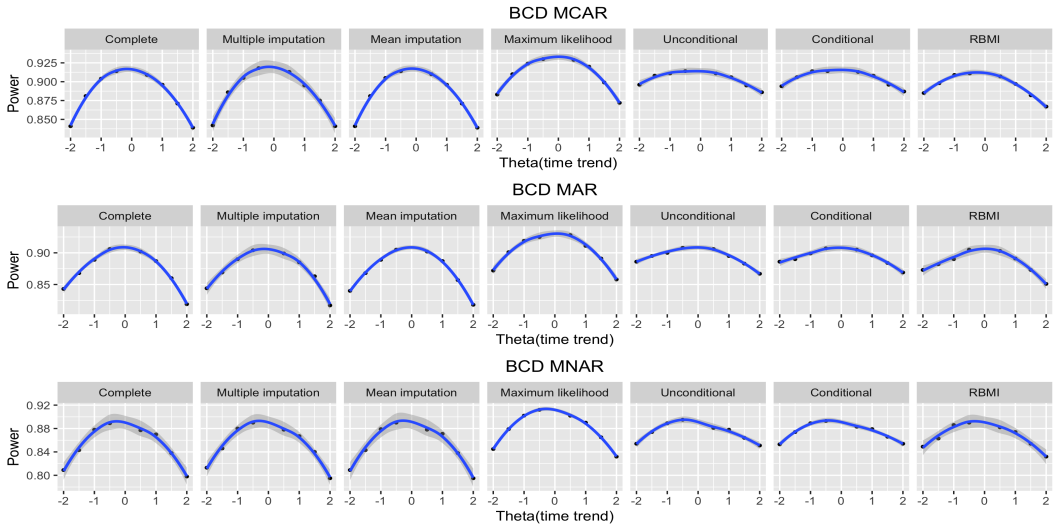

Figure 5.12: *Time trend: BCD, power ($p_{ms} = 0.1$, $n = 100$)*

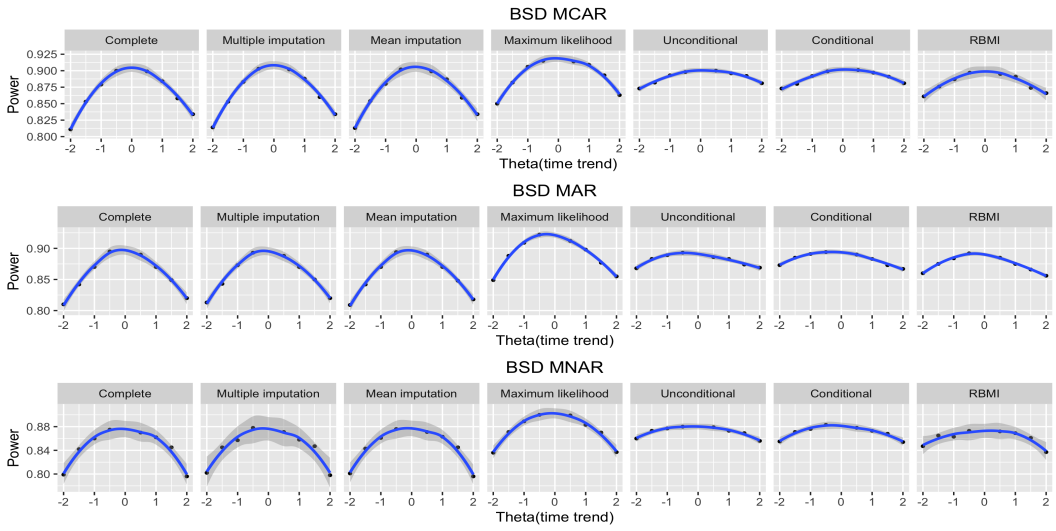

Figure 5.13: *Time trend: BSD, power ($p_{ms} = 0.1$, $n = 100$)*

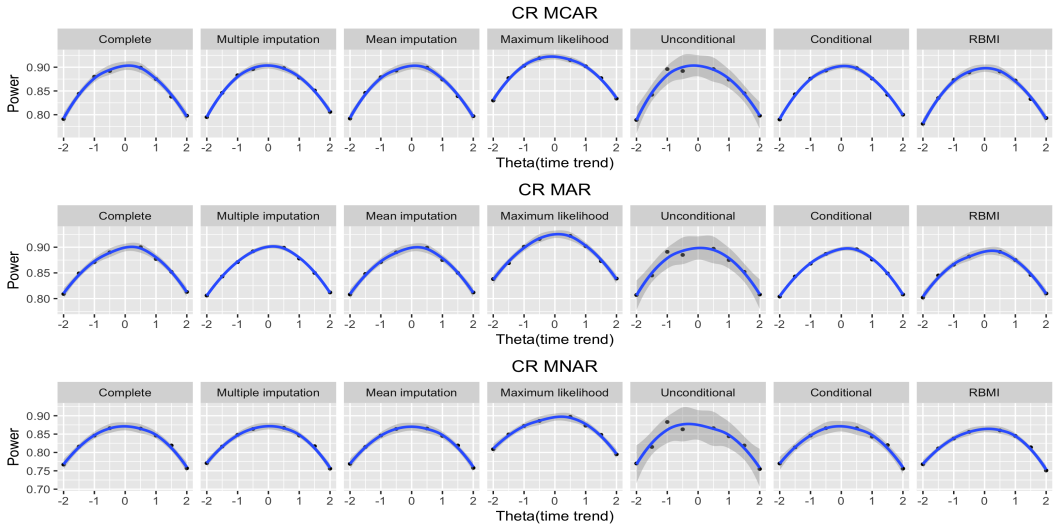

Figure 5.14: *Time trend: CR, power ($p_{ms} = 0.1$, $n = 100$)*

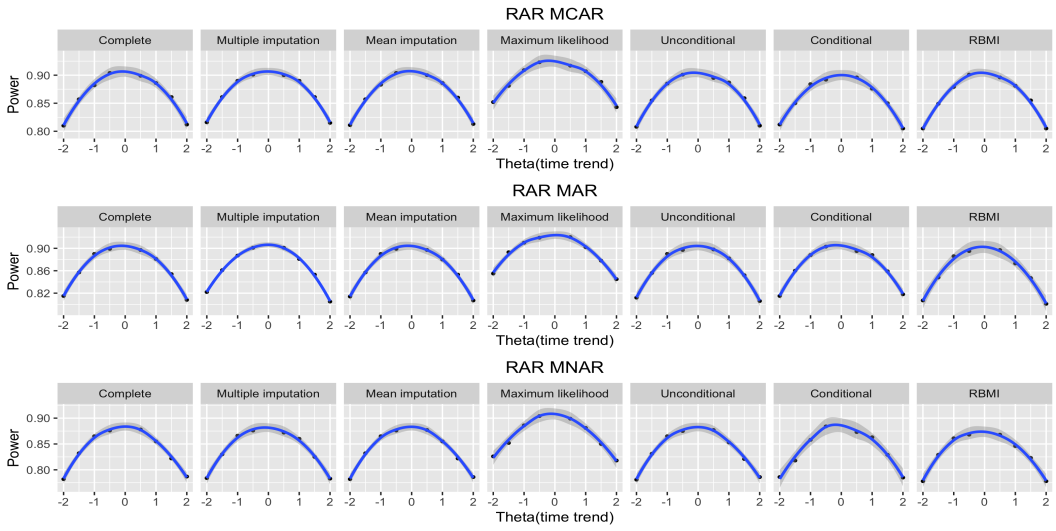

Figure 5.15: *Time trend: RAR, power ($p_{ms} = 0.1$, $n = 100$)*

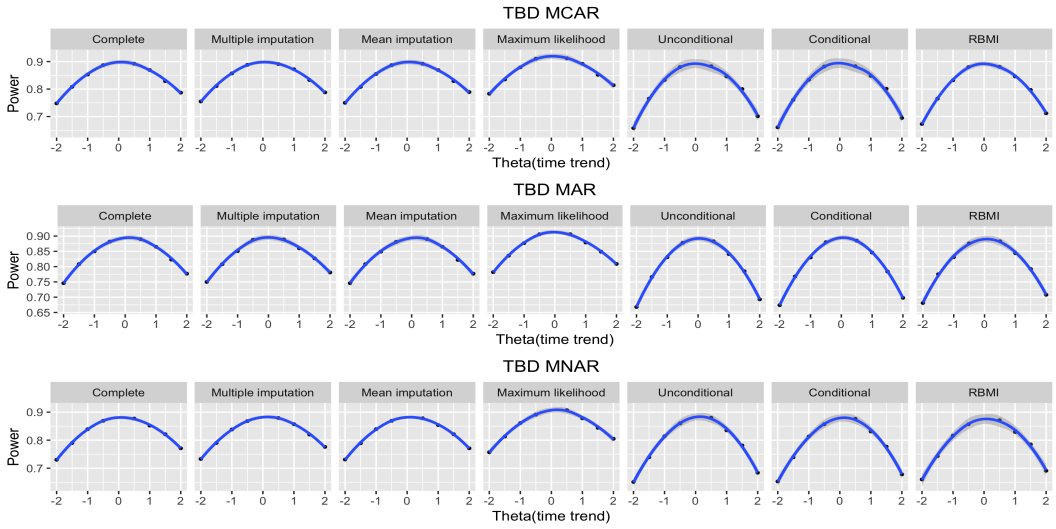

Figure 5.16: *Time trend: TBD, power ($p_{ms} = 0.1$, $n = 100$)*

5.2 Outliers in Responses

Some simulation settings for this section are different from the cases described in Chapter 3.

5.2.1 Simulation Settings

- Simulation results are separated based on the randomization procedure (BCD and RBD) and missing mechanisms (MCAR, MNAR, MAR). Details about how to simulate missing values under each missing data mechanism are available in Section 3.1. Note that the parameter r in Section 3.1 is $r = 0.7$ instead of $r = 1 : 3$ in this simulation.
- Number of simulated patients $n = 200$.
- The following are missing data methods studied, they are
  - parametric methods:
    - the complete-case analysis (complete)
    - mean imputation (mean)
    - multiple imputation with stochastic regression (MI(stochastic))
    - multiple imputation with predictive mean matching (MI(PMM))
  - Randomization-based methods:
    - the conditional reference set (conditional)
    - the unconditional reference set (unconditional)
    - the randomization-based multiple imputation (RBMI)
- The overall proportion of missingness in responses is defined as p_{ms}. If $p_{ms} = 0.1$, it means that 10% of the responses are missing, and $p_{ms} = 0.1$ or 0.2.
- Outliers in responses are simulated by Cauchy distribution. Non-outliers in responses are simulated by the normal distribution. The exact distributions used for simulating

Table 5.4: *Outilers: distribution for outliers and non-outliers in responses*

Types	Treatment	Outliers	Non-outliers
Power	A	Cauchy (Location = 0.5, scale = 1)	$N(0.5, 1)$
	B	Cauchy (Location = 0, scale = 1)	$N(0, 1)$
Type I error rate	A	Cauchy (Location = 0, scale = 1)	$N(0.5, 1)$
	B	Cauchy (Location = 0, scale = 1)	$N(0, 1)$

responses are listed in the table below.

- The proportion of outliers within responses are 0%, 5%, 10%, 15%, 20%, which are listed on the x-axis in the plots below.

5.2.2 Simulation Results

The adopted test statistic for randomization-based missing data methods is the difference in group means. Note that the test statistics for population-based approaches are t-test based statistics.

Randomization-based or population-based methods are similar in the general pattern of power and type I error rates. When the proportion of outliers increases, all methods present a decreasing power trend, and no type I error rate inflation is exhibited. It is observed that randomization-based methods are relatively less conservative compared with population-based methods.

Wang et al. (2020) also studied the type I error rates and power when responses are affected by outliers; the difference is that no missing values existed in responses. The technique used to simulate responses with outliers is also by sampling from Cauchy distribution, which is similar to the proposed setting above. In Wang et al. (2020), the adopted test statistics for randomization tests are the difference in group means, and the linear-rank test statistic. It was discovered by Wang et al. (2020) that the choice of test statistic has a

more significant impact on power; randomization tests with linear-rank statistics have larger power than the one with a difference in means under the same simulation case. When missing values and outliers coexist in responses, randomization-based missing data methods with linear-rank test statistics are conducted, and a similar fact is observed; i.e., randomization-based missing data methods are more powerful under linear-rank test statistics than the difference between group means. Data are in the Appendix.

In conclusion, when non-normal outliers and missing values coexist in patients' responses, randomization-based methods with the test statistic - differences in means - do not show a apparent advantage in power than population-based ones, and both kinds of methods present good performance in type I error rate control. To deal with such cases, selecting a more appropriate test statistic for outliers, such as the rank-based one, would be more helpful in maintaining statistical power.

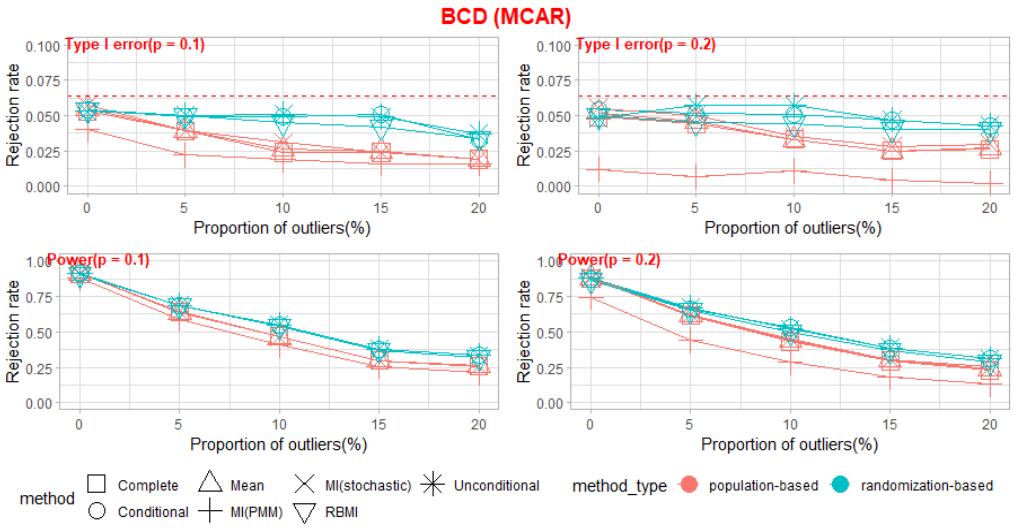

Figure 5.17: *Outliers: BCD under MCAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

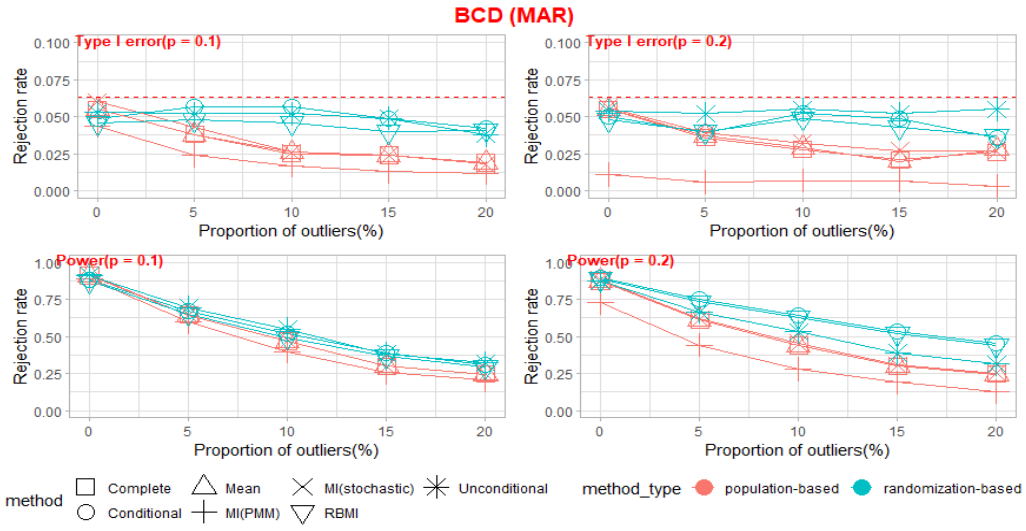

Figure 5.18: *Outliers: BCD under MAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

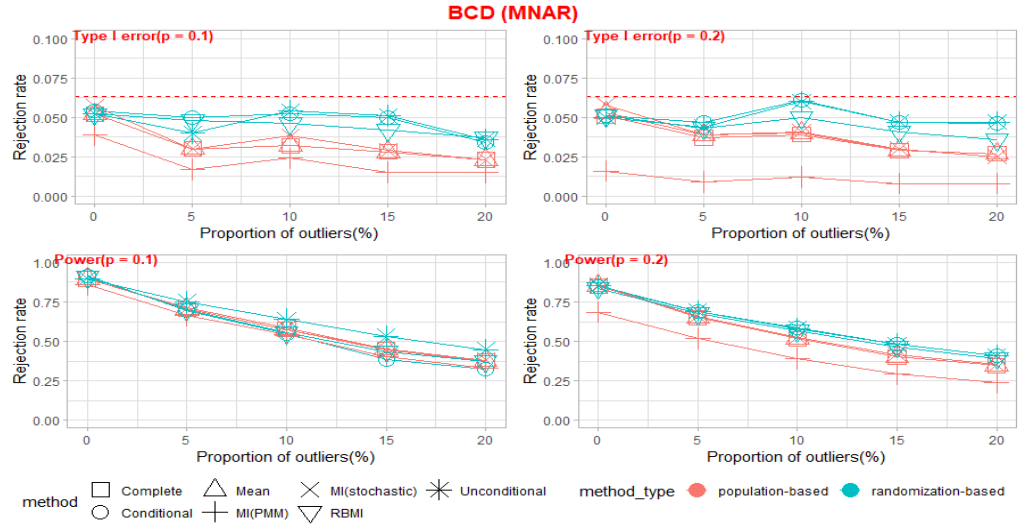

Figure 5.19: *Outliers: BCD under MNAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

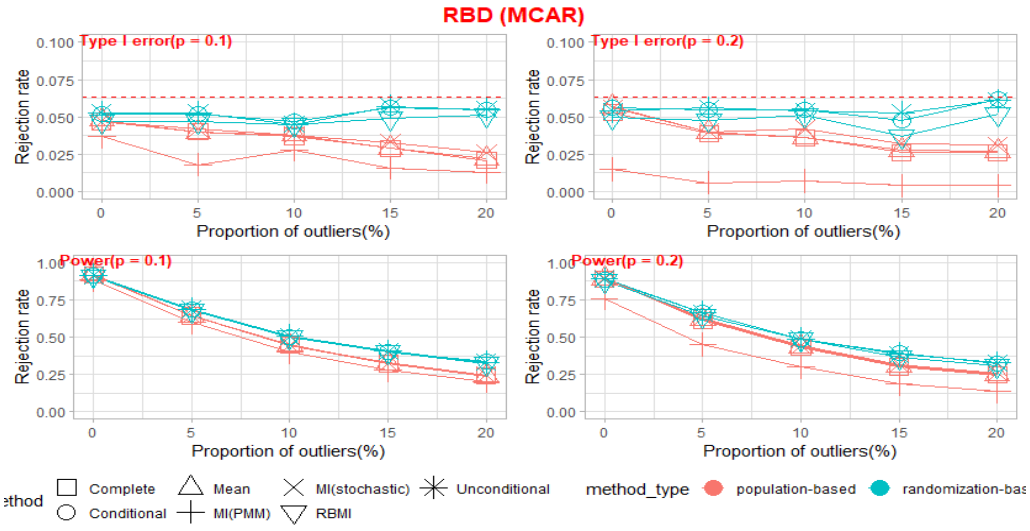

Figure 5.20: *Outliers: RBD under MCAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

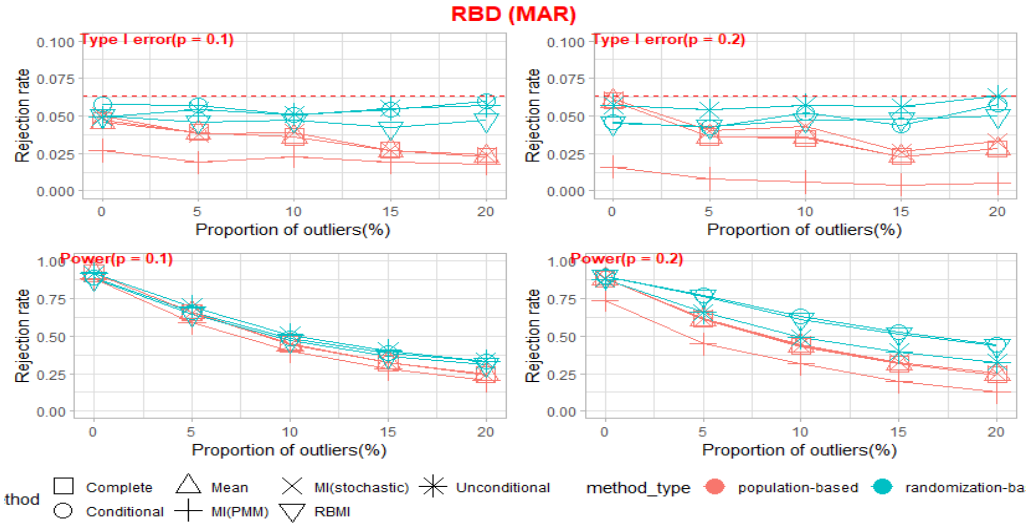

Figure 5.21: *Outliers: RBD under MAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

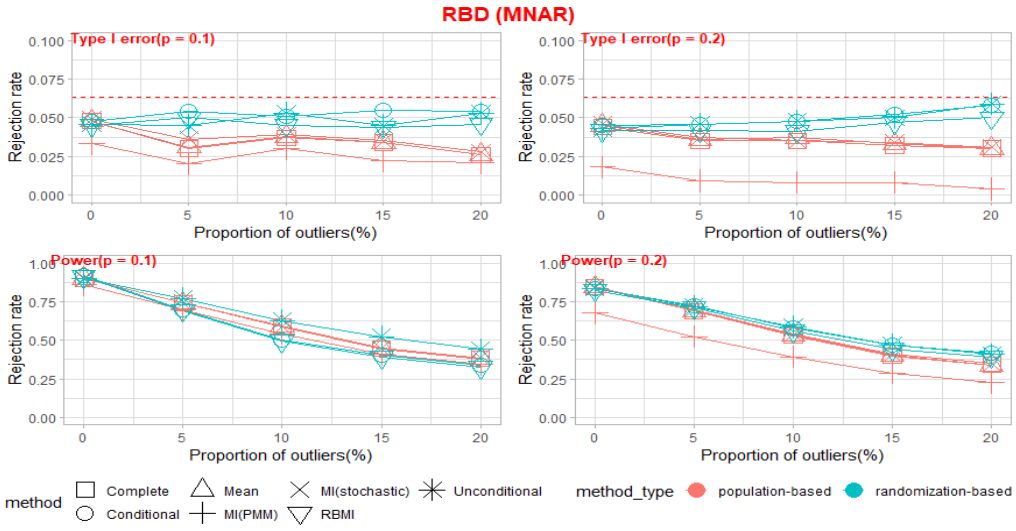

Figure 5.22: *Outliers: RBD under MNAR,* $p_{ms} = 0.1$ *(left),* $p_{ms} = 0.2$ *(right).*

5.3 Increasing Missing Proportion over Time

In previous studies, a constant overall missing proportion is assumed at each time point during a trial; i.e., the probability of a patient's responses to be missing does not change over time. In practice, we may observe a varying missing proportion over time. For instance, more patients may decide to discontinue the treatment in the later period of a trial than the early period, which creates a non-constant missing proportion in patients' responses. Based on such interest, a simulation study for evaluating randomization-based missing data methods is conducted, assuming a varying missing proportion is confronted.

5.3.1 Simulation Settings

A linear time trend is simulated in the missing proportion. Let the time trend $\theta_{gi} = g(i)$ be defined by a function that is monotone in i. A linear time trend is defined as $g(i) = i/n$.

In order to model a time trend that affects the missingness indicator, let η be the

maximum "target" proportion missingness (e.g., $\eta = 20\%$) in a trial. The missing values in responses can be simulated with a varying missing proportion η.

- Missing completely at random: $\eta_i = \eta \cdot \theta_{gi}$.
- Missing at random: $\eta_i = T_i \cdot \eta \cdot \theta_{gi} + (1 - T_i) \cdot \eta$.
- Missing not at random: $\eta_i = I(Y_i > \bar{Y}) \cdot \eta \cdot \theta_{gi} + I(Y_i \leq \bar{Y}) \cdot \eta$.

Note that η_i is defined as the probability of ith patient's response to be missing. Whether ith patient's response is missing or observed is implemented via a Bernoulli random variable with parameter η_i. Possible options for $\eta \in \{0, 0.05, 0.1, 0.15\}$ and assume $n = 200$. The distribution used to simulate responses is the Normal distribution. When calculating power, if a patient from treatments A, their responses generate from $N(0.5, 1)$; if treatment B, their responses generate from $N(0, 1)$. When calculating type I error rates, patients' responses follow $N(0, 1)$ regardless of treatment assignments.

5.3.2 Simulation Results

The test statistic for all randomization-based methods is the difference between group means. Similarly, the test statistic for population-based methods is the t-test statistic. Two randomization procedures: RBD and BCD, are used for simulating treatment assignments. Compared methods are the same as the methods in Section 5.2.1.

The simulated type I error rates and power are presented in Figures 5.23 - 5.28. Based on these figures, no significant difference between population-based methods and randomization-based methods is observed in terms of power and type I error rates. A decreasing power trend is observed when the "target" missing proportion η increases. Slight type I error rate inflation is observed under RBD and BCD. Note that they are just out of the inflation bound (indicated by the red dashed lines in plots). The conditional reference set method exhibit a relatively strong control in type I error rate regardless of randomization procedures and missing data mechanisms.

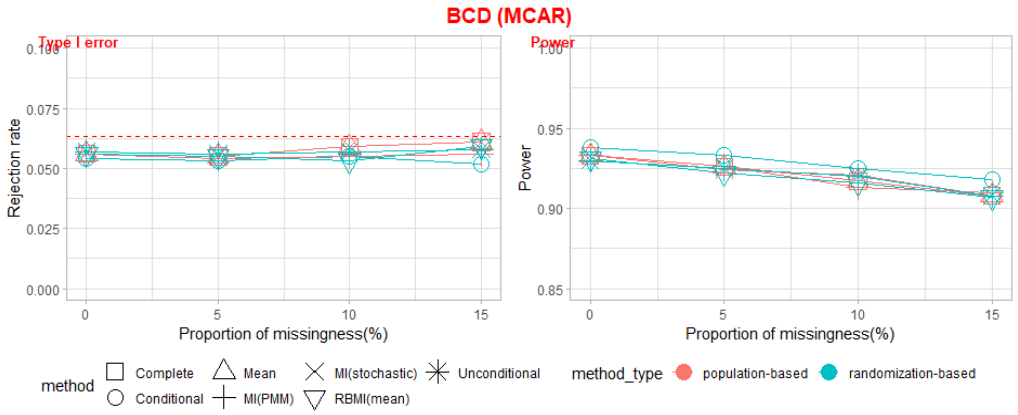

Figure 5.23: *Non-constant missing proportion: BCD under MCAR*

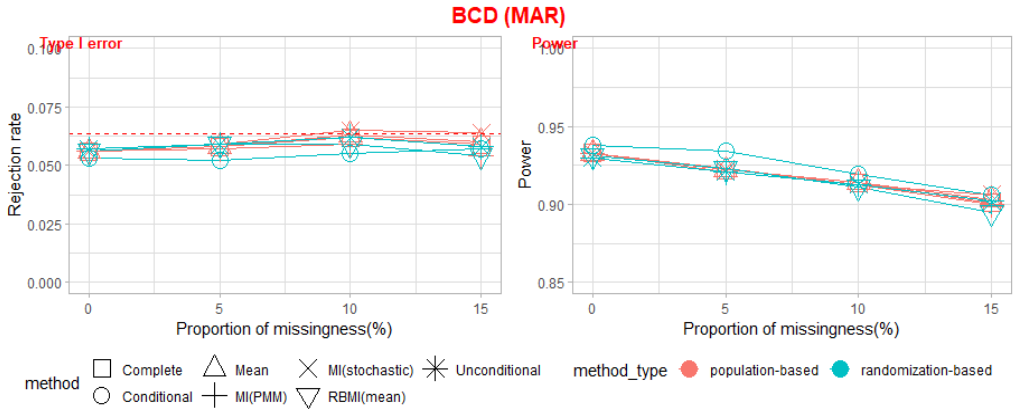

Figure 5.24: *Non-constant missing proportion: BCD under MAR*

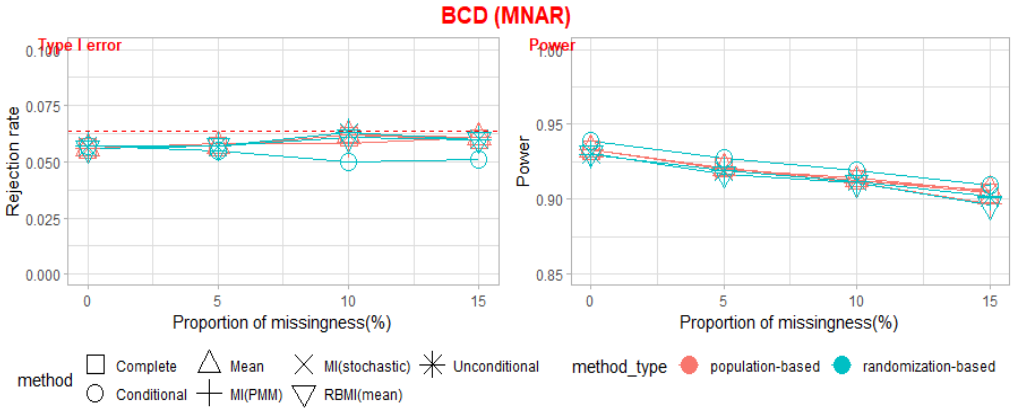

Figure 5.25: *Non-constant missing proportion: BCD under MNAR*

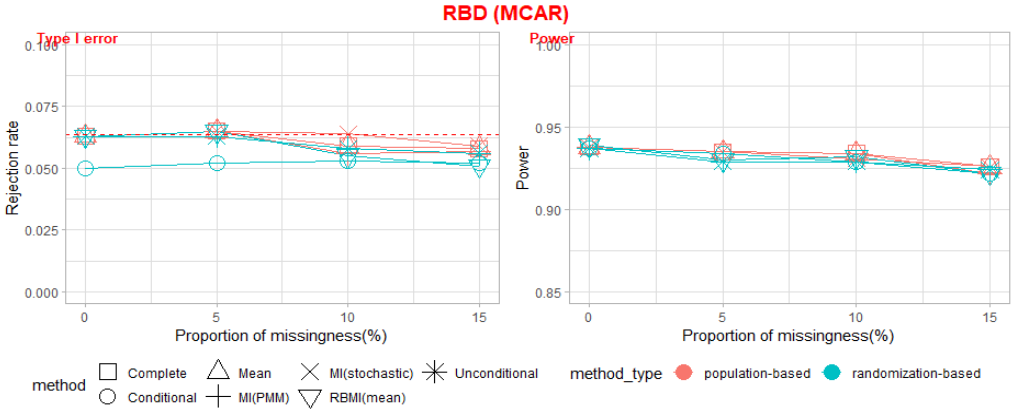

Figure 5.26: *Non-constant missing proportion: RBD under MCAR*

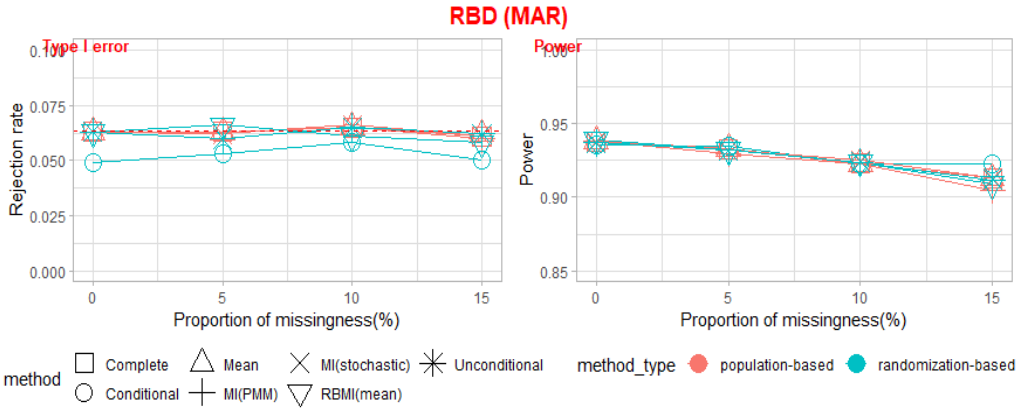

Figure 5.27: *Non-constant missing proportion: RBD under MAR*

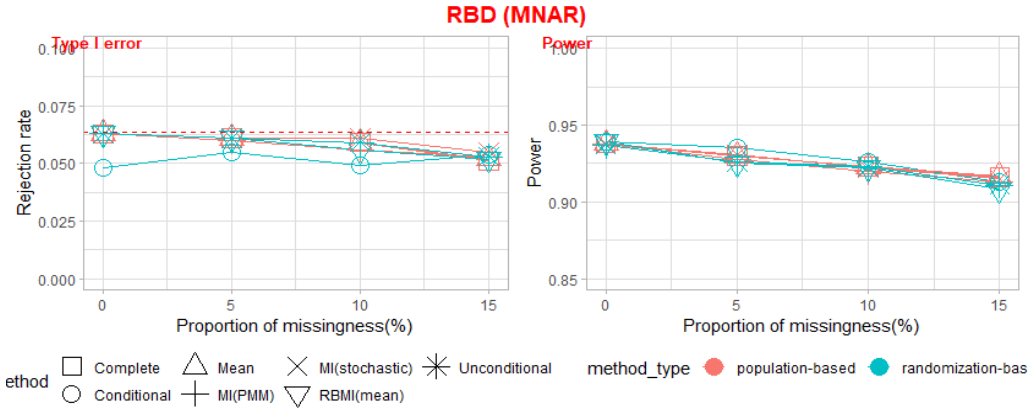

Figure 5.28: *Non-constant missing proportion: RBD under MNAR*

Chapter 6: Testing the MCAR Assumption

6.1 Randomization-based Confidence Interval

In Wang and Rosenberger (2020), randomization-based confidence intervals are designed for models using a scalar parameter to evaluate treatment effects. For instance, the study's interest is to obtain a confidence interval for a constant additive effect Δ from a two-armed randomized controlled trial.

However, the construction and interpretation of a randomization-based confidence interval differ from a population-based one. For a population-based confidence interval, data are assumed to arise from a random sample, and the confidence interval is computed according to a specific parametric distribution when constructing a confidence interval for a population parameter; e.g., the treatment effect. The interpretation for the population-based confidence interval relies on the random-sampling assumption; it is interpreted as a range covering the unknown parameter with a pre-determined probability. A population-based confidence interval can be constructed by inverting the rejection region of a corresponding hypothesis test.

The interpretation of a randomization-based confidence interval is distinct. As stated in Edgington and Onghena (2007),

> the confidence interval from randomization tests is a set of Δ values from which the hypothesis H_Δ that the treatment difference is Δ for each and every patient in the study is not rejected at the prescribed significance level based on the given set experimental data.

Based on the algorithm developed by Garthwaite (1996), Wang and Rosenberger (2020) proposed an algorithm to estimate the endpoints of a randomization-based confidence interval that does not require a random sampling assumption or a population distribution

assumption. The search algorithm developed by Garthwaite (1996) is built based on the Robbins-Monro process, and it is computationally efficient. As the number of search steps increases, the corresponding coverage probability of the estimated confidence limits is unbiased and has a smaller variance. Patients' outcomes are permuted when applying Garthwaite's method when searching for confidence limits, and treatment assignments are held fixed, which is not appropriate in some cases since randomization sequences are not always equiprobable. Moreover, Garthwaite's method only adjusts patients' responses based on the testing hypothesis once at the beginning of the search. As pointed out by Edgington and Onghena (2007),

> An equitable procedure would require modifying patient outcome data according to the hypothesis at each time a re-randomization of treatment assignments is generated.

Instead of permuting patients' responses, Wang and Rosenberger (2020)'s method updates patients' responses every time based on the newly generated sequence, and a large number of randomization sequences are generated during the searching process.

This chapter proposes a new methodology to test the missing data mechanism based on the randomization-based confidence interval developed by Wang and Rosenberger (2020). A randomization-based confidence interval derived is desired to measure the deviation of the independence assumption between missingness in responses and treatment assignments. The difference in proportions of missingness between treatments provide potential evidence to determine whether missing is completely at random or missing at random. A randomization-based confidence interval for testing the difference in missing proportion between treatments can be constructed to determine $\boldsymbol{M}$ and $\boldsymbol{T}$ are independent through the implementation of Wang and Rosenberger (2020)'s search algorithm.

6.2 Review of Randomization-based Confidence Interval

Wang and Rosenberger (2020) illustrate the idea of randomization-based confidence interval by using constant additive treatment effect as an example. The constant additive treatment effect is defined as Δ. Suppose a hypothesis H_Δ states that treatment A has an additive treatment effect Δ comparing to the treatment B, and the difference between group means as the test statistic S to estimate Δ. When iterative searching the endpoints of the confidence interval of Δ, Wang and Rosenberger (2020) 's method would generate a large number of randomization sequences, and patients' responses would be updated based on each newly generated sequence. The way to update patients' responses is illustrated at here: for each newly generated sequence, if a patient originally observed in treatment A group, is re-randomized to treatment B, the corresponding observed response would be decreased by v; and if a patient originally observed in treatment B is re-randomized to treatment A, the observed response would be increased by v. Note that v here is not fixed as the observed test statistic, the exact quantity for v when a new sequence generated in explained Section 6.3.

6.3 The Robbins-Monro Algorithm

Wang and Rosenberger (2020) discuss the process of finding the endpoints of a randomization based confidence interval with the reference to Garthwaite (1996). There are several steps for this method. Suppose the (Δ_L, Δ_U) is the interested $100(1-2\alpha)\%$ two-sided confidence interval for Δ, and $0 < \alpha < 0.5$. Let L_i, U_i be the estimates of Δ_L and Δ_U after i steps.

- First, find the starting values for the endpoints of an interval. There are M newly generated re-randomization sequences by the randomization procedure employed and each sequence is updated based on $S_{obs.}$ (i.e. $v = S_{obs.}$ in Section 6.2); and $S_{obs.}$ is the observed test statistic, for example, the observed difference in group means. Note that M is suggested as $M = (2-\alpha)/\alpha$. In practice, $M = 1000$ to provide more

precision in estimating the starting values for the endpoints of the confidence interval interested. For each newly generated randomization sequence with responses updated, a new test statistic, e.g., the difference between group means, can be calculated. For M generated sequences, M new test statistics values are available. Let t_1 be the second smallest test statistic value, and t_2 be the second largest test statistics value among these M statistics values (if $M = 1000$, t_1 be statistic value on $\alpha\%$ quantile and t_2 be the statistic value on $(1-\alpha\%)$ quantile). The starting values are for the upper limit of a confidence interval is U_0, and the starting value for the lower limit of a confidence interval is L_0 where $U_0 = S_{obs.} + (t_2 - t_1)/2$ or $L_0 = S_{obs.} - (t_2 - t_1)/2$.

- After starting values U_0 and L_0 are found, the endpoints of a confidence interval can be searched via iterative updating U_i and L_i. The following provides how to iterative search the endpoints of the upper limit and the lower limit respectively.

  - Search for the upper limit. To estimate U_{i+1}, a new treatment sequence is generated with patients' outcomes updated with U_i (i.e. $v = U_i$ in Section 6.2), and a new test statistic S_{U_i} (e.g., the difference in group means) can be obtained after responses modified by U_i. Based on a newly generated sequence with modified outcomes, the U_{i+1} is obtained by the following equation.

$$U_{i+1} = \begin{cases} U_i - c_i\alpha/i, & \text{if } S_{U_i} > S_{obs} \\ U_i + c_i(1-\alpha)/i, & \text{if } S_{U_i} \leq S_{obs} \end{cases} \tag{6.1}$$

    where $c_i = k(U_i - S_{obs.})$ and c_i is defined as a positive step length constant. The constant k is defined as $k = 2/\{z_\alpha(2\pi)^{-1/2} - exp(-z_\alpha^2/2)\}$ where z_α is the upper $100\alpha\%$ point of the standard normal distribution.

  - Search for the lower limit. To estimate L_{i+1}, a new treatment sequence is generated with patients' outcomes modified by L_i (i.e. $v = L_i$ in Section 6.2), then a new statistic S_{L_i} is obtained. Based on a newly generated sequence with modified

outcomes, the L_{i+1} is obtained by the following equation.

$$L_{i+1} = \begin{cases} L_i + c_i\alpha/i, & \text{if } S_{L_i} < S_{obs} \\ L_i - c_i(1-\alpha)/i, & \text{if } S_{U_i} \geq S_{obs} \end{cases} \tag{6.2}$$

where $c_i = k(S_{obs.} - L_i)$.

Wang and Rosenberger (2020) suggest that N steps be used when searching the endpoints of a confidence interval, and N is suggested to be a value greater than the number of sequences used for a re-randomization test. When updating U_i or L_i via formulas above, Garthwaite (1996) suggests that $m - i + 1$ should be used as the denominator rather than i to reduce the impact resulted from the rapid change in the beginning steps, where $m = \min\{50, 0.3(2-\alpha)/\alpha\}$. As pointed out by Wang and Rosenberger (2020), for most commonly encountered distributions, the previous c_i provides a convergent sequence when numerically approaching the limits. However, the previous choice of c_i does not apply to the Cauchy distribution and the two parameterizations of the exponential distribution. In general, the previous choice of c_i is recommended if no better information is available.

6.4 Implementation of Randomization-Based Confidence Intervals to Evaluate the Missing Data Mechanism

Wang and Rosenberger (2020) discuss the randomization-based confidence interval for a continuous quantity. The confidence interval for the difference in proportions of missingness between treatments is desired. To be more specific, whether a patient's response is missing or observed can be considered as a binary outcome. Based on binary outcomes created by missingness. it is of interest that an randomization-based confidence interval to determine if $\boldsymbol{M}$ and $\boldsymbol{T}$ are independent within some threshold.

If a patient's response is missing, the corresponding binary outcome equals one; otherwise, the outcome equals zero. The difference in missing proportions is equivalent to the difference in group means, if binary outcomes resulted from missingness is adopted. In Section 6.2 and 6.3, a example of finding the confidence interval of the additive treatment effect is illustrated, and the confidence interval for the difference in group means is derived according to the algorithm proposed by Wang and Rosenberger (2020). Similarly, the algorithm implemented in Section 6.3 can be applied to iterative search the confidence intervals of the difference in missing proportions based on binary outcomes resulted from missingness in responses. Then we can used the confidence interval for the difference in missing proportions derived to evaluate the independent relationship between $\boldsymbol{M}$ and $\boldsymbol{T}$. Note that the way of modifying binary outcomes is illustrated in the following example.

The quantity studied is the difference in missing proportions between treatments, it is defined as Λ and the confidence interval for Λ is desired, where Λ = the missing proportion from A - the missing proportion from B, denoted as $q_A - q_B$. The observed difference in missing proportion is defined as $\Lambda_{obs.} = q_{A,obs.} - q_{B,obs.}$. The difference in group means of patients' responses is used to estimate Λ and patients' responses are binary outcomes resulted from the missingness in patients' responses. Suppose an observed sequence of treatment assignments is (A, A, B, B) and the second and fourth patients' responses are missing (marked as the underlined treatment in the table below). During the searching process of the endpoints for a confidence interval, many new treatment assignment are generated, e.g., (A, B, B, A) (the second row in the table below). The method to update binary outcomes is the same as the one for continuous responses: if the patient observed in A is re-randomized to B, subtract v from the observed outcome; if a patient observed in B is re-randomized to A, add v to the observed outcome. Note that the modification to the binary responses should not only happen in locations that have missingness, but also every treatment within a sequence. The whole process of iterative searching the endpoints for a confidence interval is the same illustrated in Section 6.3, except that patients' responses are binary outcomes resulted from the missingness. The algorithm developed by Wang and

Rosenberger (2020) can be applied to seeking the randomization-based confidence interval of the difference in missing proportion between treatments. An optimal target is to find the maximum imbalance between missing proportions that will not span the zero as a threshold with a pre-defined confidence level. If the imbalance observed is greater than the threshold, it concludes that the independence between $\boldsymbol{M}$ and $\boldsymbol{T}$ is violated under the predetermined confidence level.

Table 6.1: *Methods of modifying binary responses based on missingness*

Treatment assignments	Outcomes data based on missingness	Outcomes values (binary)	Difference in missing proportions ($q_A - q_B$)	Differencee in missing proportions ($q_A - q_B$) under $v = \Lambda_{obs}$
$A, \underline{A}, B, \underline{B}$	m_1, m_2, m_3, m_4	$(0, 1, 0, 1)$	$(m_1 + m_2)/2 - (m_3 + m_4)/2$	Λ_{obs} = (0+1)/2 - (0+1)/2 = 0
$A, \underline{B}, B, \underline{A}$	$m_1, m_2 - v, m_3, m_4 + v$	$(0, 1 - v, 0, 1 + v)$	$(m_1 + m_4)/2 - (m_2 + m_3)/2$	$(0 + 1 + \Lambda_{obs})/2 - (1 - \Lambda_{obs} + 0)/2 = 0$
$A, \underline{B}, A, \underline{B}$	$m_1, m_2 - v, m_3 + v, m_4$	$(0, 1 - v, v, 1)$	$(m_1 + m_3)/2 - (m_2 + m_4)/2$	$(0 + \Lambda_{obs})/2 - (1 - \Lambda_{obs} + 1)/2 = -1$
...	...	...	...	...

6.5 Examples on Randomization-based Confidence Intervals

Several examples are simulated for illustration purposes, and the quantity studied is the difference in missing proportions between A and B. Suppose the sample size $n = 200$, and the randomization procedures implemented are the biased coin design (BCD) or randomized block design (RBD). The population-based confidence intervals, i.e., the Z intervals derived from the two-sample proportion test, are calculated for comparison. Table 6.2 below shows that the randomization-based confidence intervals derived are similar to the corresponding Z intervals.

Table 6.2: Examples for randomization-based confidence intervals

Designs	$q_{A,obs.}$	$q_{B,obs.}$	observed difference in missing proportions (in %)	80% confidence interval (population-based)	80% confidence interval (randomization-based)
RBD	1%	21%	(1% - 21% = -20%)	(-0.2537, -0.1463)	(-0.2554, -0.1443)
RBD	3%	12%	(3% - 12% = -9%)	(-0.1370, -0.0430)	(-0.1373, -0.0413)
RBD	3%	6%	(3% - 6% = -3%)	(-0.0675, 0.0075)	(-0.0676, 0.0006)
BCD	1%	23%	(1% - 23% = -22%)	(-0.2754, -0.1646)	(-0.2764, -0.1634)
BCD	1%	11%	(1% - 11% = -10%)	(-0.1421, -0.0579)	(-0.1431, -0.0585)
BCD	1%	6%	(1% - 6% = -5%)	(-0.0830, -0.0170)	(-0.0833, -0.0192)

When different missing proportions are observed, one interesting scientific question is how large the difference in missing proportions is required to conclude that the two missing proportions are distinct with certain pre-determined confidence. A universal threshold that measures the deviation from the independence assumption between $\boldsymbol{M}$ and $\boldsymbol{T}$ is desired. Once a significant violation of the unconditional method's assumption (see Section 2.2 for details) is confirmed, the conditional reference set method is preferred. This interest is investigated by constructing randomization-based confidence intervals of different imbalance values in missing proportions.

Examples below illustrate the idea of the desired threshold, and simulation results are derived from 80% two-sided confidence intervals. Suppose the smaller missing proportion is observed under treatment B and the number of patients is 200. The question becomes ascertaining the required minimum imbalance in missing proportions such that the corresponding two-sided confidence interval of $q_A - q_B$ does not cover zero. The following simulation results are obtained by grid search, and the step in changing the missing proportion is 1%. Note that adopting finer steps in missing proportions could provide more

accurate values in searching the required minimum imbalances. Table 6.3 contains simulation results from two randomization procedures: BCD and RBD. The first column is the observed smaller missing proportion $q_{B,obs.}$ by assumption. The second and third columns are the required minimum imbalance in missing proportions to state that q_A is significantly larger than q_B. Under if the observed smaller missing proportion $q_{B,obs.}$ is 3%, then the missing proportion $q_{A,obs.}$ must be greater than or equal to $3\% + 5\% = 8\%$ to conclude that the missing proportion in A is significantly larger than the one from B.

Table 6.3: *Threshold for stating non-equivalence in missing proportions (80% confidence)*

$q_{B,obs.}$	minimum imbalance (BCD)	minimum imbalance (RBD)
0.01	0.04	0.02
0.02	0.05	0.05
0.03	0.05	0.04
0.04	0.05	0.05
0.05	0.05	0.05
0.06	0.06	0.05
0.07	0.05	0.04
0.08	0.06	0.05
0.09	0.06	0.06
0.10	0.07	0.06
0.11	0.07	0.06
0.12	0.07	0.06
0.13	0.07	0.07
0.14	0.08	0.07
0.15	0.08	0.07
0.16	0.08	0.08
0.17	0.08	0.08
0.18	0.08	0.07
0.19	0.09	0.08
0.20	0.09	0.07
0.21	0.09	0.07
0.22	0.09	0.08
0.23	0.09	0.08
0.24	0.09	0.08
0.25	0.09	0.08

Chapter 7: Conclusion and Future work

7.1 Conclusion

Population-based methods which require the random sampling assumption have been widely used to evaluate the treatment effect in randomized clinical trials. However, the random sampling assumption is invalid for patients in a randomized clinical trial due to the uniqueness of an experimental design. Patients recruited in a trial essentially form a population, which weakens the generality of the statistical inference based on the random sampling assumption. The randomization test, which has been neglected for many years, exhibits its advantages in delivering reliable inference results without requiring the random sampling assumption. The randomization test is built upon the randomization procedure that is applied in the trial. Randomization procedures themselves provide the basis for obtaining a more valid and objective treatment effect evaluation.

Missing data is an inevitable problem in clinical trials. Population-based solutions for missing values are extensively researched. The lack of randomization-based missing data methods significantly hinders the application of randomization tests in practice.

The contribution of this thesis is that it proposes and evaluates randomization-based missing data methods by comparing them with population-based missing data methods under different simulation scenarios. This comparison is conducted from two perspectives depending on whether homogeneity affects responses.

In general, population-based methods are comparable to population-based methods regarding power and type I error rates when evaluating treatment effects. To achieve a fair comparison, when responses are continuous, the test-statistic for randomization-based methods is the difference between group means; the test-statistic for population-based methods is the t-test. For binary responses, the test-statistic for randomization-based methods is

rank-based test statistic (i.e., a randomization-based analogue of the χ^2-test); the test for population methods is the Pearson's χ^2-test. Multiple simulation scenarios are generated by varying the missing data mechanisms, sample sizes, and overall missing proportion.

Under homogeneous responses, regardless of the type of responses or the involving randomization procedures, randomization-based missing methods show similar performance to the population-based approaches in most simulation scenarios; the "worst" method, the "best-worst" method, are not recommended due to their poor performance in power and type I error rates. The maximum likelihood is not recommended for continuous responses because of the potential type I error rate inflation. Note that the "worst" and "best-worst" methods are excluded from the simulation for responses with heterogeneity because of their poor performance.

When heterogeneity exists in continuous responses, randomization-based missing data methods are evaluated from three perspectives. When responses are affected by time trends, randomization-based methods present advantages in preserving type I error rates, especially for procedures such as the truncated binomial design, which tends to be highly affected because of the design itself. The performance of missing data methods in type I error rates is more design-specified. Block-based designs exhibit advantages in type I error rates control over non-block designs. The conclusions above are obtained from simulations involving eight previously discussed randomization procedures in Chapter 1. The other two aspects evaluated for heterogeneity are responses with outliers, and the non-constant missing proportion over time. When responses are affected by outliers, randomization-based methods and population-based methods are similar in power and type I error rates. In previous simulations, a constant missing proportion over time is assumed. However, the probability of observing missingess in responses change over time. Randomization-based missing data methods are assessed based on this interest, and a linearly increasing missing proportion over time is simulated. No apparent difference in power and type I error rates are observed between randomization-based and population-based methods. The conditional reference set method shows a slightly smaller type I error rate than the remaining methods.

In Chapter 6, a randomization-based confidence interval of the difference between missing proportions is proposed to measure the deviation of the independence assumption between the missingness in responses and treatment assignments. The proposed interval is built on the Wang and Rosenberger (2020)'s randomization-based confidence interval, where their interval is designed for a quantity (e.g., the additive treatment effect) on the real line. In this thesis, the quantity interested ranges from [-1,1], i.e., the difference in missing proportions between treatments. Wang and Rosenberger (2020)'s method has been modified to deal with the difference in missing proportions. The randomization-based confidence intervals for the difference in missing proportions are compared with the Z intervals obtained from the two-sample proportion Z test through simulations; similar confidence intervals are observed under the same observed missing proportions. A threshold that can measure the deviation of the independence assumption between $\boldsymbol{M}$ and $\boldsymbol{T}$ is desired. To be more specific, the threshold is constructed by searching the minimum imbalance between missing proportions such that an 80% randomization-based confidence interval derived for the difference in missing proportion does not cover zero. In practice, to clarify the violation of independence assumption, it is recommended to calculate a randomization-based confidence interval for the difference in missing proportions with a user-selected confidence level. If the randomization-based confidence interval calculated does not cover zero, then the conditional reference set method should be adopted.

7.2 Future Work

An R package containing these R modules will be built for open source use to the public. One limit of the current implementation of the conditional reference set method is that it is computationally burdensome to sample enough sequences in the conditional reference set, when the number of missing observations between treatments differs substantially. This is because sequences in the conditional reference set are sampled from the unconditional reference set, and the sequences do not satisfy the requirement are discarded. The current solution to mitigate such weakness is to implement the conditional reference set method via

a language with higher computational speed such as C++.

In clinical trials, generalized linear mixed models (GLMM) are commonly used in analyzing longitudinal data, for example, repeated measures on biomarkers from patients at multiple time points when comparing two treatment effects. Parhat et al. (2014) extended the application of randomization tests on GLMM by providing the corresponding algorithm in the context of a two-armed clinical trial. Particularly, Parhat et al. (2014) investigate the treatment effect variation over the repeated measures such as whether a treatment has a time-varying effect for a patient. However, no missing data method is available when applying a randomization test under GLMM. Thus, it is worth investigating corresponding randomization-based missing data methods. Other missing data methods for longitudinal studies will be employed in simulation studies for comparison, such as multiple imputation, the last observation carried forward method when different missing data mechanisms are assumed. Also, when applying the randomization-based methods, further investigation is needed to study why the type I error rate is inflated under specific randomization procedures, e.g., BSD, which is also observed in Wang et al. (2020).

Appendix A: An Appendix

A.1 Results under Homogeneity (Continuous Responses)

A.1.1 Type I Error Rates

A.1.1.1 MCAR

Table A.1: *Homogeneous responses (continuous): MCAR, type I error rates, n=50,* p_{ms} *=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.076	0.036	0.044	0.043	0.044	0.060	0.040	0.043	0.040
2	BSD	0.102	0.049	0.058	0.061	0.058	0.066	0.056	0.055	0.055
3	CR	0.087	0.043	0.049	0.049	0.049	0.057	0.050	0.055	0.049
4	PBD (blocksize = 4)	0.090	0.061	0.052	0.053	0.052	0.061	0.055	0.049	0.055
5	PBD (blocksize = 6)	0.081	0.049	0.045	0.045	0.044	0.061	0.046	0.045	0.042
6	RAR	0.071	0.047	0.052	0.051	0.052	0.061	0.049	0.042	0.051
7	RBD	0.082	0.040	0.040	0.039	0.040	0.048	0.042	0.046	0.039
8	TBD	0.077	0.040	0.060	0.060	0.060	0.070	0.060	0.060	0.059

Table A.2: *Homogeneous responses (continuous): MCAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.195	0.035	0.040	0.040	0.040	0.064	0.038	0.039	0.037
2	BSD	0.222	0.053	0.058	0.061	0.058	0.079	0.059	0.058	0.058
3	CR	0.223	0.050	0.053	0.048	0.052	0.073	0.051	0.044	0.049
4	PBD (blocksize = 4)	0.197	0.057	0.053	0.058	0.054	0.077	0.059	0.044	0.056
5	PBD (blocksize = 6)	0.197	0.047	0.046	0.047	0.046	0.065	0.045	0.046	0.046
6	RAR	0.187	0.052	0.048	0.047	0.048	0.071	0.047	0.042	0.049
7	RBD	0.208	0.031	0.042	0.041	0.042	0.052	0.044	0.050	0.040
8	TBD	0.197	0.037	0.058	0.057	0.059	0.075	0.062	0.057	0.056

Table A.3: *Homogeneous responses (continuous): MCAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.138	0.040	0.051	0.053	0.051	0.056	0.052	0.049	0.051
2	BSD	0.141	0.061	0.069	0.070	0.069	0.073	0.067	0.047	0.067
3	CR	0.136	0.042	0.053	0.052	0.053	0.059	0.052	0.051	0.052
4	PBD (blocksize = 4)	0.141	0.052	0.055	0.057	0.055	0.062	0.057	0.050	0.057
5	PBD (blocksize = 6)	0.141	0.041	0.047	0.046	0.047	0.053	0.042	0.051	0.041
6	RAR	0.132	0.060	0.065	0.067	0.065	0.075	0.065	0.062	0.062
7	RBD	0.114	0.037	0.041	0.043	0.041	0.046	0.042	0.045	0.039
8	TBD	0.139	0.060	0.055	0.054	0.055	0.064	0.055	0.057	0.050

Table A.4: *Homogeneous responses (continuous): MCAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.367	0.033	0.046	0.050	0.046	0.058	0.046	0.048	0.047
2	BSD	0.376	0.064	0.062	0.063	0.062	0.077	0.061	0.048	0.061
3	CR	0.373	0.051	0.052	0.052	0.052	0.059	0.051	0.059	0.050
4	PBD (blocksize = 4)	0.376	0.055	0.057	0.061	0.058	0.074	0.059	0.050	0.056
5	PBD (blocksize = 6)	0.393	0.044	0.049	0.050	0.049	0.065	0.051	0.050	0.043
6	RAR	0.367	0.058	0.058	0.061	0.058	0.084	0.058	0.054	0.057
7	RBD	0.360	0.039	0.047	0.049	0.048	0.058	0.049	0.045	0.045
8	TBD	0.369	0.052	0.055	0.061	0.055	0.074	0.056	0.061	0.056

A.1.1.2 MAR

Table A.5: *Homogeneous responses (continuous): MAR, type I error rates, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.078	0.048	0.044	0.046	0.044	0.057	0.041	0.041	0.039
2	BSD	0.102	0.058	0.060	0.059	0.061	0.067	0.054	0.054	0.056
3	CR	0.098	0.063	0.054	0.057	0.054	0.062	0.055	0.046	0.054
4	PBD (blocksize = 4)	0.100	0.067	0.049	0.051	0.049	0.060	0.056	0.047	0.054
5	PBD (blocksize = 6)	0.079	0.055	0.049	0.044	0.047	0.061	0.044	0.045	0.044
6	RAR	0.074	0.060	0.055	0.053	0.054	0.065	0.053	0.042	0.054
7	RBD	0.086	0.041	0.037	0.037	0.038	0.047	0.038	0.045	0.039
8	TBD	0.081	0.045	0.062	0.063	0.062	0.071	0.060	0.055	0.056

Table A.6: *Homogeneous responses (continuous): MAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.213	0.082	0.043	0.045	0.043	0.065	0.046	0.036	0.045
2	BSD	0.219	0.100	0.060	0.059	0.060	0.077	0.062	0.051	0.057
3	CR	0.218	0.091	0.046	0.048	0.046	0.065	0.045	0.047	0.043
4	PBD (blocksize = 4)	0.205	0.090	0.058	0.055	0.058	0.073	0.058	0.049	0.058
5	PBD (blocksize = 6)	0.201	0.088	0.048	0.049	0.047	0.071	0.044	0.046	0.046
6	RAR	0.179	0.080	0.057	0.058	0.056	0.080	0.056	0.043	0.058
7	RBD	0.195	0.048	0.037	0.036	0.036	0.056	0.040	0.044	0.035
8	TBD	0.198	0.079	0.057	0.060	0.058	0.073	0.057	0.050	0.057

Table A.7: *Homogeneous responses (continuous): MAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.140	0.060	0.045	0.048	0.045	0.055	0.047	0.049	0.044
2	BSD	0.147	0.080	0.064	0.065	0.064	0.077	0.066	0.048	0.066
3	CR	0.143	0.081	0.050	0.051	0.050	0.058	0.051	0.053	0.051
4	PBD (blocksize = 4)	0.133	0.077	0.058	0.057	0.058	0.068	0.060	0.048	0.059
5	PBD (blocksize = 6)	0.134	0.067	0.043	0.041	0.043	0.048	0.044	0.057	0.042
6	RAR	0.136	0.079	0.059	0.060	0.059	0.068	0.059	0.060	0.059
7	RBD	0.116	0.065	0.040	0.040	0.040	0.045	0.041	0.046	0.040
8	TBD	0.144	0.080	0.057	0.057	0.057	0.062	0.055	0.059	0.054

Table A.8: *Homogeneous responses (continuous): MAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.382	0.149	0.049	0.053	0.048	0.064	0.048	0.055	0.051
2	BSD	0.396	0.146	0.064	0.069	0.063	0.085	0.066	0.056	0.061
3	CR	0.384	0.143	0.052	0.050	0.052	0.065	0.052	0.052	0.050
4	PBD (blocksize = 4)	0.404	0.140	0.055	0.055	0.055	0.069	0.054	0.047	0.053
5	PBD (blocksize = 6)	0.403	0.137	0.046	0.047	0.046	0.061	0.044	0.051	0.041
6	RAR	0.384	0.152	0.063	0.067	0.063	0.079	0.064	0.059	0.063
7	RBD	0.374	0.066	0.043	0.042	0.043	0.054	0.042	0.041	0.039
8	TBD	0.381	0.136	0.061	0.062	0.060	0.081	0.062	0.058	0.054

A.1.1.3 MNAR

Table A.9: *Homogeneous responses (continuous): MNAR, type I error rates, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.078	0.042	0.037	0.037	0.037	0.049	0.036	0.045	0.037
2	BSD	0.090	0.063	0.052	0.053	0.052	0.066	0.054	0.045	0.055
3	CR	0.075	0.046	0.048	0.050	0.048	0.056	0.048	0.041	0.050
4	PBD (blocksize = 4)	0.085	0.052	0.063	0.062	0.063	0.072	0.060	0.039	0.060
5	PBD (blocksize = 6)	0.084	0.050	0.043	0.042	0.043	0.052	0.044	0.050	0.043
6	RAR	0.083	0.061	0.052	0.054	0.052	0.063	0.051	0.048	0.051
7	RBD	0.086	0.055	0.044	0.045	0.045	0.052	0.039	0.050	0.040
8	TBD	0.077	0.044	0.056	0.055	0.056	0.075	0.059	0.047	0.058

Table A.10: *Homogeneous responses (continuous): MNAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.189	0.045	0.039	0.040	0.041	0.060	0.039	0.029	0.038
2	BSD	0.200	0.041	0.055	0.055	0.055	0.072	0.054	0.040	0.054
3	CR	0.183	0.041	0.047	0.049	0.048	0.064	0.045	0.051	0.046
4	PBD (blocksize = 4)	0.200	0.047	0.058	0.061	0.057	0.075	0.054	0.040	0.056
5	PBD (blocksize = 6)	0.201	0.057	0.049	0.049	0.049	0.059	0.047	0.048	0.046
6	RAR	0.205	0.047	0.055	0.056	0.055	0.074	0.054	0.043	0.056
7	RBD	0.194	0.052	0.040	0.046	0.041	0.063	0.040	0.046	0.042
8	TBD	0.193	0.049	0.054	0.057	0.054	0.077	0.058	0.049	0.052

Table A.11: *Homogeneous responses (continuous): MNAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.126	0.035	0.043	0.043	0.043	0.050	0.042	0.048	0.041
2	BSD	0.128	0.057	0.065	0.063	0.065	0.070	0.063	0.053	0.064
3	CR	0.118	0.033	0.053	0.052	0.052	0.059	0.050	0.051	0.050
4	PBD (blocksize = 4)	0.133	0.060	0.053	0.055	0.053	0.059	0.054	0.046	0.055
5	PBD (blocksize = 6)	0.128	0.046	0.043	0.042	0.043	0.051	0.045	0.056	0.042
6	RAR	0.128	0.044	0.055	0.055	0.055	0.064	0.057	0.060	0.054
7	RBD	0.119	0.047	0.044	0.043	0.044	0.050	0.043	0.047	0.044
8	TBD	0.118	0.047	0.053	0.056	0.053	0.062	0.056	0.054	0.054

Table A.12: *Homogeneous responses (continuous): MNAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.357	0.042	0.037	0.037	0.037	0.051	0.038	0.035	0.035
2	BSD	0.359	0.048	0.060	0.059	0.060	0.076	0.064	0.055	0.061
3	CR	0.362	0.045	0.049	0.049	0.049	0.066	0.047	0.055	0.046
4	PBD (blocksize = 4)	0.370	0.049	0.058	0.059	0.058	0.078	0.063	0.051	0.059
5	PBD (blocksize = 6)	0.378	0.052	0.049	0.049	0.049	0.066	0.046	0.059	0.045
6	RAR	0.353	0.050	0.063	0.066	0.063	0.080	0.064	0.051	0.064
7	RBD	0.354	0.046	0.042	0.046	0.043	0.057	0.043	0.043	0.044
8	TBD	0.365	0.040	0.057	0.056	0.057	0.067	0.055	0.062	0.056

A.1.2 Power

A.1.2.1 MCAR

Table A.13: *Homogeneous responses (continuous): MCAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.659	0.864	0.936	0.935	0.936	0.949	0.936	0.933	0.933
2	BSD	0.628	0.815	0.901	0.905	0.901	0.936	0.901	0.881	0.900
3	CR	0.630	0.839	0.906	0.907	0.906	0.916	0.906	0.924	0.904
4	PBD (blocksize = 4)	0.649	0.855	0.927	0.925	0.927	0.938	0.922	0.917	0.920
5	PBD (blocksize = 6)	0.651	0.848	0.923	0.919	0.923	0.932	0.923	0.904	0.918
6	RAR	0.644	0.857	0.918	0.921	0.918	0.935	0.920	0.923	0.919
7	RBD	0.622	0.844	0.920	0.919	0.920	0.931	0.911	0.918	0.914
8	TBD	0.647	0.862	0.919	0.917	0.919	0.933	0.921	0.931	0.903

Table A.14: *Homogeneous responses (continuous): MCAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.304	0.767	0.920	0.920	0.920	0.941	0.922	0.923	0.920
2	BSD	0.272	0.720	0.880	0.881	0.881	0.924	0.878	0.898	0.873
3	CR	0.276	0.748	0.891	0.894	0.890	0.908	0.889	0.919	0.885
4	PBD (blocksize = 4)	0.273	0.745	0.908	0.913	0.909	0.936	0.911	0.899	0.904
5	PBD (blocksize = 6)	0.270	0.749	0.909	0.907	0.910	0.932	0.910	0.894	0.902
6	RAR	0.273	0.768	0.906	0.904	0.906	0.930	0.907	0.910	0.905
7	RBD	0.277	0.734	0.891	0.892	0.891	0.920	0.882	0.905	0.887
8	TBD	0.289	0.760	0.897	0.898	0.899	0.925	0.904	0.921	0.880

Table A.15: *Homogeneous responses (continuous): MCAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.560	0.862	0.927	0.928	0.928	0.938	0.930	0.921	0.927
2	BSD	0.559	0.845	0.914	0.910	0.914	0.925	0.915	0.905	0.916
3	CR	0.553	0.848	0.909	0.910	0.911	0.919	0.911	0.919	0.906
4	PBD (blocksize = 4)	0.545	0.846	0.916	0.913	0.916	0.924	0.915	0.910	0.917
5	PBD (blocksize = 6)	0.561	0.856	0.920	0.921	0.921	0.934	0.916	0.906	0.912
6	RAR	0.568	0.857	0.918	0.920	0.918	0.924	0.917	0.911	0.916
7	RBD	0.569	0.880	0.927	0.924	0.927	0.938	0.928	0.925	0.929
8	TBD	0.574	0.847	0.919	0.917	0.919	0.922	0.913	0.913	0.896

Table A.16: *Homogeneous responses (continuous): MCAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.171	0.766	0.916	0.919	0.916	0.936	0.916	0.905	0.912
2	BSD	0.183	0.758	0.902	0.902	0.902	0.919	0.898	0.887	0.896
3	CR	0.169	0.759	0.896	0.901	0.898	0.926	0.901	0.909	0.895
4	PBD (blocksize = 4)	0.161	0.756	0.904	0.902	0.904	0.917	0.898	0.903	0.897
5	PBD (blocksize = 6)	0.163	0.774	0.907	0.910	0.907	0.927	0.902	0.892	0.902
6	RAR	0.178	0.775	0.900	0.907	0.902	0.927	0.903	0.899	0.901
7	RBD	0.172	0.781	0.911	0.911	0.911	0.930	0.910	0.910	0.905
8	TBD	0.175	0.762	0.901	0.904	0.901	0.915	0.899	0.897	0.878

A.1.2.2 MAR

Table A.17: *Homogeneous responses (continuous): MAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.647	0.775	0.939	0.940	0.939	0.948	0.939	0.936	0.939
2	BSD	0.621	0.727	0.901	0.901	0.901	0.933	0.901	0.886	0.899
3	CR	0.626	0.730	0.909	0.909	0.909	0.921	0.906	0.934	0.909
4	PBD (blocksize = 4)	0.641	0.751	0.922	0.924	0.922	0.941	0.922	0.913	0.923
5	PBD (blocksize = 6)	0.627	0.748	0.921	0.918	0.921	0.933	0.918	0.910	0.912
6	RAR	0.646	0.770	0.922	0.918	0.922	0.935	0.921	0.921	0.918
7	RBD	0.620	0.739	0.913	0.914	0.913	0.927	0.909	0.913	0.910
8	TBD	0.638	0.757	0.914	0.912	0.914	0.928	0.914	0.928	0.900

Table A.18: *Homogeneous responses (continuous): MAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.313	0.527	0.915	0.915	0.915	0.937	0.915	0.920	0.908
2	BSD	0.253	0.494	0.882	0.884	0.882	0.924	0.887	0.900	0.884
3	CR	0.281	0.503	0.893	0.894	0.893	0.917	0.891	0.916	0.888
4	PBD (blocksize = 4)	0.279	0.504	0.897	0.904	0.897	0.933	0.899	0.906	0.901
5	PBD (blocksize = 6)	0.276	0.499	0.896	0.899	0.897	0.923	0.893	0.890	0.891
6	RAR	0.277	0.519	0.898	0.899	0.899	0.922	0.899	0.904	0.892
7	RBD	0.268	0.494	0.902	0.899	0.901	0.928	0.898	0.912	0.898
8	TBD	0.293	0.533	0.903	0.902	0.903	0.931	0.907	0.915	0.882

Table A.19: *Homogeneous responses (continuous): MAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.561	0.724	0.928	0.927	0.927	0.938	0.929	0.922	0.928
2	BSD	0.541	0.714	0.911	0.911	0.911	0.934	0.910	0.901	0.906
3	CR	0.550	0.718	0.911	0.909	0.910	0.923	0.911	0.922	0.910
4	PBD (blocksize = 4)	0.530	0.714	0.921	0.919	0.921	0.929	0.923	0.916	0.921
5	PBD (blocksize = 6)	0.552	0.732	0.927	0.928	0.927	0.936	0.925	0.911	0.920
6	RAR	0.571	0.735	0.920	0.924	0.920	0.931	0.921	0.915	0.920
7	RBD	0.566	0.746	0.926	0.926	0.926	0.938	0.930	0.927	0.927
8	TBD	0.561	0.712	0.914	0.912	0.913	0.920	0.910	0.913	0.888

Table A.20: *Homogeneous responses (continuous): MAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.180	0.425	0.911	0.905	0.910	0.929	0.908	0.905	0.907
2	BSD	0.182	0.435	0.900	0.895	0.899	0.905	0.893	0.877	0.891
3	CR	0.154	0.430	0.903	0.901	0.902	0.925	0.902	0.902	0.895
4	PBD (blocksize = 4)	0.159	0.406	0.907	0.907	0.906	0.925	0.906	0.904	0.904
5	PBD (blocksize = 6)	0.169	0.416	0.917	0.914	0.917	0.930	0.910	0.898	0.906
6	RAR	0.186	0.449	0.903	0.907	0.903	0.920	0.903	0.905	0.900
7	RBD	0.157	0.447	0.914	0.909	0.914	0.932	0.911	0.919	0.911
8	TBD	0.171	0.433	0.896	0.893	0.896	0.916	0.891	0.890	0.863

A.1.2.3 MNAR

Table A.21: *Homogeneous responses (continuous): MNAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.628	0.759	0.921	0.920	0.920	0.943	0.922	0.919	0.921
2	BSD	0.592	0.744	0.899	0.897	0.898	0.924	0.894	0.879	0.892
3	CR	0.647	0.786	0.907	0.911	0.909	0.929	0.911	0.919	0.909
4	PBD (blocksize = 4)	0.641	0.774	0.918	0.917	0.917	0.932	0.921	0.912	0.915
5	PBD (blocksize = 6)	0.625	0.761	0.909	0.908	0.910	0.929	0.914	0.909	0.906
6	RAR	0.613	0.756	0.908	0.910	0.908	0.919	0.908	0.914	0.907
7	RBD	0.621	0.761	0.906	0.911	0.906	0.923	0.908	0.907	0.909
8	TBD	0.633	0.768	0.908	0.904	0.908	0.922	0.904	0.919	0.892

Table A.22: *Homogeneous responses (continuous): MNAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.284	0.572	0.898	0.900	0.898	0.931	0.897	0.899	0.892
2	BSD	0.268	0.557	0.877	0.879	0.880	0.905	0.872	0.896	0.870
3	CR	0.288	0.577	0.880	0.885	0.877	0.909	0.878	0.896	0.876
4	PBD (blocksize = 4)	0.287	0.555	0.900	0.901	0.900	0.926	0.885	0.889	0.879
5	PBD (blocksize = 6)	0.292	0.553	0.891	0.894	0.888	0.920	0.888	0.879	0.882
6	RAR	0.282	0.558	0.882	0.888	0.883	0.913	0.883	0.889	0.887
7	RBD	0.277	0.572	0.885	0.887	0.884	0.914	0.884	0.890	0.881
8	TBD	0.272	0.550	0.870	0.873	0.872	0.895	0.867	0.895	0.843

Table A.23: *Homogeneous responses (continuous): MNAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.575	0.787	0.924	0.924	0.924	0.933	0.925	0.910	0.923
2	BSD	0.575	0.768	0.895	0.897	0.895	0.907	0.894	0.895	0.891
3	CR	0.558	0.764	0.904	0.903	0.904	0.917	0.901	0.900	0.897
4	PBD (blocksize = 4)	0.541	0.761	0.913	0.913	0.913	0.922	0.903	0.916	0.902
5	PBD (blocksize = 6)	0.551	0.771	0.909	0.909	0.909	0.923	0.907	0.902	0.907
6	RAR	0.567	0.771	0.907	0.906	0.907	0.920	0.905	0.913	0.905
7	RBD	0.558	0.793	0.919	0.921	0.919	0.927	0.917	0.922	0.916
8	TBD	0.553	0.782	0.911	0.909	0.911	0.922	0.912	0.911	0.903

Table A.24: *Homogeneous responses (continuous): MNAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Multiple imputation	Mean imputation	Maximum likelihood	Unconditional	Conditional	RBMI
1	BCD	0.142	0.591	0.893	0.891	0.893	0.910	0.891	0.886	0.891
2	BSD	0.159	0.581	0.876	0.876	0.877	0.899	0.880	0.873	0.875
3	CR	0.156	0.559	0.872	0.876	0.872	0.897	0.873	0.875	0.870
4	PBD (blocksize = 4)	0.167	0.573	0.882	0.878	0.882	0.905	0.876	0.899	0.877
5	PBD (blocksize = 6)	0.181	0.576	0.890	0.891	0.890	0.916	0.886	0.884	0.883
6	RAR	0.179	0.601	0.888	0.888	0.887	0.911	0.883	0.887	0.875
7	RBD	0.159	0.593	0.894	0.892	0.894	0.920	0.889	0.898	0.885
8	TBD	0.165	0.579	0.886	0.886	0.886	0.908	0.886	0.884	0.848

A.2 Results under Homogeneity (Binary Responses)

A.2.1 Type I Error Rates

A.2.1.1 MCAR

Table A.25: *Homogeneous responses (binary): MCAR, type I error rates, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.060	0.033	0.030	0.033	0.036	0.069	0.068
2	BSD	0.039	0.035	0.034	0.039	0.039	0.041	0.038
3	CR	0.063	0.041	0.030	0.032	0.047	0.047	0.043
4	PBD (blocksize = 4)	0.059	0.040	0.034	0.033	0.049	0.048	0.048
5	PBD (blocksize = 6)	0.061	0.037	0.036	0.033	0.043	0.029	0.040
6	RAR	0.038	0.022	0.032	0.025	0.041	0.028	0.030
7	RBD	0.049	0.027	0.030	0.027	0.032	0.030	0.030
8	TBD	0.064	0.033	0.030	0.032	0.040	0.040	0.034

Table A.26: *Homogeneous responses (binary): MCAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.107	0.034	0.032	0.047	0.039	0.071	0.065
2	BSD	0.092	0.030	0.029	0.041	0.038	0.042	0.037
3	CR	0.123	0.045	0.032	0.052	0.048	0.049	0.050
4	PBD (blocksize = 4)	0.118	0.036	0.036	0.047	0.049	0.042	0.042
5	PBD (blocksize = 6)	0.120	0.037	0.036	0.049	0.053	0.033	0.041
6	RAR	0.093	0.020	0.029	0.045	0.046	0.032	0.032
7	RBD	0.086	0.024	0.027	0.041	0.034	0.031	0.025
8	TBD	0.107	0.036	0.031	0.049	0.041	0.040	0.032

Table A.27: *Homogeneous responses (binary): MCAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.076	0.043	0.040	0.045	0.046	0.075	0.068
2	BSD	0.082	0.036	0.032	0.051	0.041	0.038	0.043
3	CR	0.105	0.063	0.046	0.066	0.066	0.063	0.064
4	PBD (blocksize = 4)	0.075	0.034	0.037	0.046	0.049	0.041	0.033
5	PBD (blocksize = 6)	0.086	0.050	0.042	0.059	0.049	0.033	0.049
6	RAR	0.080	0.043	0.040	0.053	0.051	0.049	0.047
7	RBD	0.056	0.023	0.028	0.035	0.033	0.030	0.025
8	TBD	0.092	0.044	0.035	0.049	0.045	0.040	0.042

Table A.28: *Homogeneous responses (binary): MCAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.198	0.040	0.038	0.053	0.050	0.072	0.065
2	BSD	0.190	0.038	0.032	0.050	0.041	0.041	0.039
3	CR	0.235	0.057	0.041	0.086	0.060	0.059	0.058
4	PBD (blocksize = 4)	0.160	0.025	0.040	0.052	0.050	0.038	0.030
5	PBD (blocksize = 6)	0.188	0.045	0.039	0.058	0.050	0.032	0.047
6	RAR	0.179	0.043	0.046	0.060	0.054	0.043	0.047
7	RBD	0.167	0.019	0.027	0.044	0.032	0.029	0.023
8	TBD	0.220	0.050	0.041	0.058	0.049	0.040	0.045

A.2.1.2 MAR

Table A.29: *Homogeneous responses (binary): MAR, type I error rates, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.040	0.032	0.029	0.030	0.039	0.065	0.065
2	BSD	0.037	0.028	0.029	0.031	0.038	0.037	0.038
3	CR	0.051	0.038	0.029	0.036	0.044	0.045	0.047
4	PBD (blocksize = 4)	0.052	0.040	0.039	0.036	0.051	0.049	0.046
5	PBD (blocksize = 6)	0.046	0.041	0.034	0.030	0.049	0.030	0.042
6	RAR	0.028	0.024	0.029	0.021	0.038	0.029	0.033
7	RBD	0.033	0.028	0.030	0.028	0.042	0.035	0.030
8	TBD	0.046	0.037	0.033	0.036	0.039	0.039	0.034

Table A.30: *Homogeneous responses (binary): MAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.070	0.031	0.031	0.045	0.045	0.073	0.067
2	BSD	0.062	0.031	0.033	0.042	0.046	0.044	0.039
3	CR	0.084	0.043	0.029	0.051	0.051	0.049	0.048
4	PBD (blocksize = 4)	0.078	0.037	0.035	0.050	0.050	0.045	0.046
5	PBD (blocksize = 6)	0.081	0.039	0.030	0.048	0.047	0.033	0.038
6	RAR	0.055	0.023	0.029	0.031	0.046	0.033	0.031
7	RBD	0.064	0.025	0.029	0.035	0.043	0.034	0.029
8	TBD	0.084	0.043	0.032	0.048	0.049	0.044	0.038

Table A.31: *Homogeneous responses (binary): MAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.060	0.046	0.040	0.057	0.052	0.072	0.065
2	BSD	0.055	0.038	0.029	0.045	0.037	0.037	0.036
3	CR	0.084	0.060	0.042	0.069	0.059	0.063	0.060
4	PBD (blocksize = 4)	0.052	0.037	0.038	0.043	0.053	0.040	0.033
5	PBD (blocksize = 6)	0.063	0.053	0.042	0.050	0.052	0.028	0.048
6	RAR	0.054	0.044	0.044	0.050	0.051	0.050	0.044
7	RBD	0.046	0.024	0.028	0.030	0.032	0.032	0.024
8	TBD	0.067	0.041	0.036	0.054	0.043	0.038	0.040

Table A.32: *Homogeneous responses (binary): MAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.120	0.046	0.041	0.064	0.051	0.074	0.067
2	BSD	0.120	0.046	0.028	0.054	0.042	0.041	0.037
3	CR	0.150	0.066	0.045	0.076	0.061	0.065	0.063
4	PBD (blocksize = 4)	0.101	0.033	0.038	0.048	0.049	0.042	0.030
5	PBD (blocksize = 6)	0.118	0.051	0.044	0.057	0.056	0.036	0.051
6	RAR	0.118	0.048	0.044	0.053	0.057	0.045	0.043
7	RBD	0.107	0.031	0.022	0.030	0.035	0.028	0.021
8	TBD	0.128	0.058	0.040	0.061	0.050	0.044	0.045

A.2.1.3 MNAR

Table A.33: *Homogeneous responses (binary): MNAR, type I error rates, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.056	0.034	0.027	0.031	0.040	0.075	0.070
2	BSD	0.048	0.029	0.031	0.034	0.038	0.041	0.040
3	CR	0.064	0.045	0.027	0.035	0.046	0.045	0.046
4	PBD (blocksize = 4)	0.056	0.036	0.028	0.027	0.036	0.041	0.037
5	PBD (blocksize = 6)	0.057	0.037	0.032	0.028	0.043	0.027	0.037
6	RAR	0.035	0.020	0.024	0.020	0.031	0.030	0.025
7	RBD	0.043	0.024	0.027	0.023	0.034	0.029	0.028
8	TBD	0.058	0.035	0.029	0.033	0.041	0.038	0.031

Table A.34: *Homogeneous responses (binary): MNAR, type I error rates, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.104	0.035	0.026	0.046	0.035	0.069	0.059
2	BSD	0.111	0.035	0.028	0.047	0.035	0.042	0.036
3	CR	0.130	0.033	0.024	0.043	0.039	0.036	0.036
4	PBD (blocksize = 4)	0.122	0.032	0.035	0.040	0.044	0.046	0.040
5	PBD (blocksize = 6)	0.122	0.035	0.032	0.040	0.044	0.032	0.035
6	RAR	0.105	0.017	0.026	0.035	0.038	0.034	0.027
7	RBD	0.106	0.033	0.032	0.037	0.038	0.029	0.026
8	TBD	0.120	0.037	0.030	0.037	0.040	0.037	0.036

Table A.35: *Homogeneous responses (binary): MNAR, type I error rates, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.078	0.039	0.036	0.051	0.044	0.066	0.070
2	BSD	0.085	0.035	0.033	0.045	0.040	0.040	0.038
3	CR	0.107	0.058	0.039	0.058	0.064	0.063	0.063
4	PBD (blocksize = 4)	0.071	0.024	0.034	0.046	0.044	0.033	0.028
5	PBD (blocksize = 6)	0.076	0.045	0.041	0.049	0.052	0.026	0.048
6	RAR	0.077	0.044	0.044	0.052	0.053	0.046	0.049
7	RBD	0.064	0.018	0.027	0.031	0.033	0.029	0.022
8	TBD	0.094	0.050	0.042	0.055	0.052	0.037	0.051

Table A.36: *Homogeneous responses (binary): MNAR, type I error rates, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.222	0.037	0.033	0.063	0.049	0.073	0.059
2	BSD	0.216	0.039	0.035	0.048	0.047	0.045	0.043
3	CR	0.229	0.066	0.045	0.077	0.063	0.063	0.063
4	PBD (blocksize = 4)	0.165	0.018	0.027	0.043	0.044	0.025	0.019
5	PBD (blocksize = 6)	0.203	0.038	0.043	0.057	0.049	0.030	0.048
6	RAR	0.200	0.035	0.043	0.063	0.053	0.039	0.043
7	RBD	0.197	0.024	0.028	0.050	0.038	0.032	0.030
8	TBD	0.225	0.046	0.035	0.070	0.042	0.037	0.037

A.2.2 Power

A.2.2.1 MCAR

Table A.37: *Homogeneous responses (binary): MCAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.748	0.867	0.877	0.871	0.886	0.926	0.922
2	BSD	0.738	0.856	0.871	0.880	0.879	0.883	0.880
3	CR	0.754	0.870	0.872	0.873	0.893	0.896	0.894
4	PBD (blocksize = 4)	0.727	0.859	0.879	0.884	0.886	0.877	0.869
5	PBD (blocksize = 6)	0.743	0.866	0.882	0.878	0.889	0.881	0.884
6	RAR	0.733	0.854	0.877	0.877	0.885	0.891	0.881
7	RBD	0.717	0.847	0.871	0.869	0.885	0.874	0.861
8	TBD	0.749	0.861	0.870	0.885	0.877	0.885	0.873

Table A.38: *Homogeneous responses (binary): MCAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.548	0.841	0.851	0.861	0.876	0.913	0.909
2	BSD	0.532	0.824	0.837	0.855	0.854	0.860	0.858
3	CR	0.579	0.841	0.834	0.869	0.867	0.867	0.866
4	PBD (blocksize = 4)	0.537	0.827	0.846	0.862	0.870	0.867	0.857
5	PBD (blocksize = 6)	0.535	0.830	0.837	0.871	0.874	0.866	0.860
6	RAR	0.529	0.803	0.846	0.860	0.879	0.872	0.860
7	RBD	0.529	0.803	0.840	0.852	0.874	0.857	0.848
8	TBD	0.491	0.829	0.846	0.870	0.859	0.868	0.858

Table A.39: *Homogeneous responses (binary): MCAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.670	0.862	0.860	0.855	0.878	0.903	0.907
2	BSD	0.664	0.849	0.836	0.848	0.856	0.857	0.857
3	CR	0.694	0.879	0.861	0.869	0.879	0.878	0.883
4	PBD (blocksize = 4)	0.622	0.809	0.837	0.842	0.854	0.843	0.826
5	PBD (blocksize = 6)	0.671	0.840	0.849	0.856	0.858	0.860	0.851
6	RAR	0.628	0.811	0.843	0.841	0.858	0.871	0.850
7	RBD	0.634	0.820	0.841	0.833	0.861	0.852	0.839
8	TBD	0.673	0.851	0.859	0.857	0.872	0.861	0.864

Table A.40: *Homogeneous responses (binary): MCAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.390	0.829	0.841	0.841	0.858	0.897	0.889
2	BSD	0.399	0.811	0.813	0.842	0.832	0.834	0.830
3	CR	0.406	0.830	0.835	0.858	0.856	0.858	0.856
4	PBD (blocksize = 4)	0.345	0.770	0.824	0.832	0.845	0.828	0.803
5	PBD (blocksize = 6)	0.416	0.809	0.833	0.841	0.849	0.846	0.835
6	RAR	0.376	0.794	0.827	0.823	0.845	0.853	0.832
7	RBD	0.210	0.797	0.827	0.844	0.842	0.831	0.820
8	TBD	0.240	0.820	0.840	0.856	0.852	0.841	0.843

A.2.2.2 MAR

Table A.41: *Homogeneous responses (binary): MAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.768	0.819	0.869	0.866	0.885	0.926	0.926
2	BSD	0.757	0.815	0.859	0.863	0.869	0.877	0.874
3	CR	0.766	0.836	0.870	0.875	0.889	0.892	0.890
4	PBD (blocksize = 4)	0.757	0.824	0.876	0.875	0.883	0.883	0.871
5	PBD (blocksize = 6)	0.766	0.824	0.886	0.870	0.893	0.883	0.885
6	RAR	0.746	0.804	0.880	0.875	0.887	0.893	0.879
7	RBD	0.734	0.802	0.880	0.877	0.891	0.888	0.870
8	TBD	0.770	0.825	0.870	0.868	0.872	0.884	0.868

Table A.42: *Homogeneous responses (binary): MAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.582	0.739	0.851	0.875	0.879	0.911	0.909
2	BSD	0.585	0.740	0.833	0.864	0.849	0.860	0.855
3	CR	0.604	0.768	0.837	0.869	0.870	0.874	0.878
4	PBD (blocksize = 4)	0.547	0.729	0.858	0.871	0.873	0.862	0.862
5	PBD (blocksize = 6)	0.558	0.728	0.846	0.864	0.873	0.863	0.860
6	RAR	0.546	0.698	0.851	0.866	0.877	0.874	0.864
7	RBD	0.546	0.698	0.843	0.865	0.869	0.852	0.841
8	TBD	0.503	0.736	0.846	0.875	0.864	0.872	0.860

Table A.43: *Homogeneous responses (binary): MAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.714	0.804	0.871	0.873	0.884	0.910	0.911
2	BSD	0.698	0.794	0.839	0.861	0.859	0.859	0.855
3	CR	0.721	0.830	0.865	0.881	0.882	0.881	0.879
4	PBD (blocksize = 4)	0.655	0.764	0.846	0.858	0.861	0.849	0.833
5	PBD (blocksize = 6)	0.711	0.799	0.855	0.859	0.872	0.860	0.860
6	RAR	0.680	0.767	0.839	0.856	0.859	0.874	0.849
7	RBD	0.678	0.771	0.842	0.850	0.855	0.852	0.842
8	TBD	0.708	0.798	0.854	0.874	0.869	0.857	0.862

Table A.44: *Homogeneous responses (binary): MAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.477	0.706	0.851	0.855	0.869	0.893	0.893
2	BSD	0.482	0.701	0.828	0.856	0.848	0.847	0.844
3	CR	0.495	0.717	0.841	0.855	0.862	0.864	0.867
4	PBD (blocksize = 4)	0.416	0.651	0.821	0.830	0.841	0.824	0.805
5	PBD (blocksize = 6)	0.480	0.701	0.825	0.844	0.845	0.848	0.827
6	RAR	0.436	0.667	0.824	0.842	0.848	0.850	0.832
7	RBD	0.232	0.667	0.822	0.842	0.839	0.827	0.815
8	TBD	0.306	0.694	0.849	0.860	0.863	0.846	0.850

A.2.2.3 MNAR

Table A.45: *Homogeneous responses (binary): MNAR, power, n=50, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.756	0.846	0.871	0.866	0.887	0.924	0.924
2	BSD	0.745	0.847	0.862	0.875	0.873	0.875	0.872
3	CR	0.751	0.849	0.859	0.860	0.884	0.886	0.886
4	PBD (blocksize = 4)	0.738	0.845	0.875	0.871	0.878	0.874	0.871
5	PBD (blocksize = 6)	0.741	0.849	0.876	0.870	0.887	0.879	0.877
6	RAR	0.725	0.820	0.876	0.867	0.883	0.891	0.880
7	RBD	0.718	0.822	0.869	0.870	0.885	0.872	0.864
8	TBD	0.759	0.843	0.876	0.876	0.881	0.881	0.875

Table A.46: *Homogeneous responses (binary): MNAR, power, n=50, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.536	0.787	0.852	0.842	0.875	0.912	0.910
2	BSD	0.544	0.796	0.828	0.844	0.845	0.853	0.851
3	CR	0.572	0.811	0.833	0.859	0.866	0.869	0.868
4	PBD (blocksize = 4)	0.520	0.772	0.831	0.850	0.859	0.853	0.850
5	PBD (blocksize = 6)	0.536	0.778	0.843	0.860	0.869	0.845	0.857
6	RAR	0.512	0.744	0.835	0.837	0.863	0.857	0.845
7	RBD	0.512	0.744	0.816	0.839	0.844	0.833	0.824
8	TBD	0.468	0.778	0.836	0.866	0.849	0.854	0.844

Table A.47: *Homogeneous responses (binary): MNAR, power, n=100, p_{ms}=0.05*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.671	0.833	0.853	0.853	0.867	0.904	0.903
2	BSD	0.653	0.821	0.829	0.852	0.850	0.853	0.852
3	CR	0.676	0.846	0.851	0.870	0.869	0.870	0.870
4	PBD (blocksize = 4)	0.621	0.783	0.835	0.831	0.850	0.833	0.821
5	PBD (blocksize = 6)	0.670	0.818	0.840	0.845	0.853	0.851	0.842
6	RAR	0.635	0.804	0.840	0.835	0.856	0.857	0.848
7	RBD	0.634	0.807	0.829	0.832	0.842	0.833	0.826
8	TBD	0.669	0.825	0.851	0.860	0.862	0.853	0.857

Table A.48: *Homogeneous responses (binary): MNAR, power, n=100, p_{ms}=0.1*

	Randomization	Best & Worst	Worst	Complete-case	Logistic	Unconditional	Conditional	RBMI
1	BCD	0.396	0.772	0.821	0.829	0.840	0.877	0.876
2	BSD	0.401	0.770	0.801	0.818	0.824	0.826	0.823
3	CR	0.406	0.771	0.810	0.837	0.842	0.844	0.842
4	PBD (blocksize = 4)	0.345	0.718	0.804	0.819	0.826	0.801	0.785
5	PBD (blocksize = 6)	0.410	0.759	0.806	0.829	0.823	0.811	0.809
6	RAR	0.344	0.737	0.807	0.812	0.825	0.832	0.816
7	RBD	0.192	0.733	0.799	0.806	0.821	0.802	0.797
8	TBD	0.237	0.754	0.823	0.852	0.838	0.822	0.826

A.3 Results under Heterogeneity (Time Trends)

A.3.1 Power

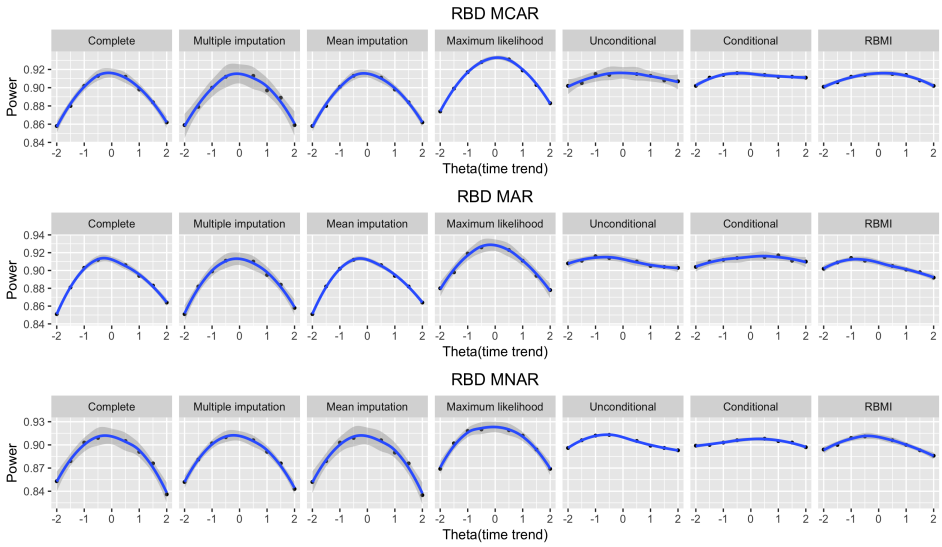

Figure A.1: *Time trend: RBD, power,* $p_{ms} = 0.05$*,* $n = 50$

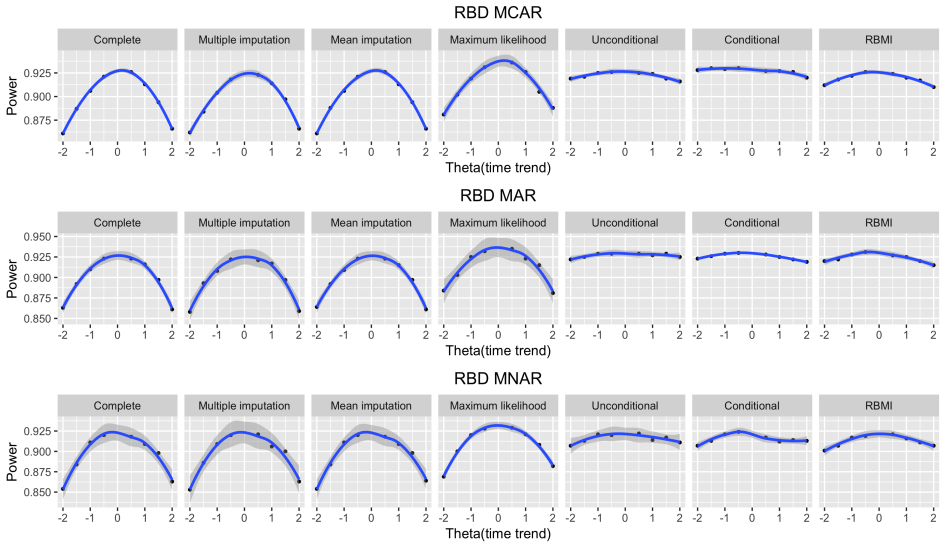

Figure A.2: *Time trend: RBD, power,* $p_{ms} = 0.05$*,* $n = 100$

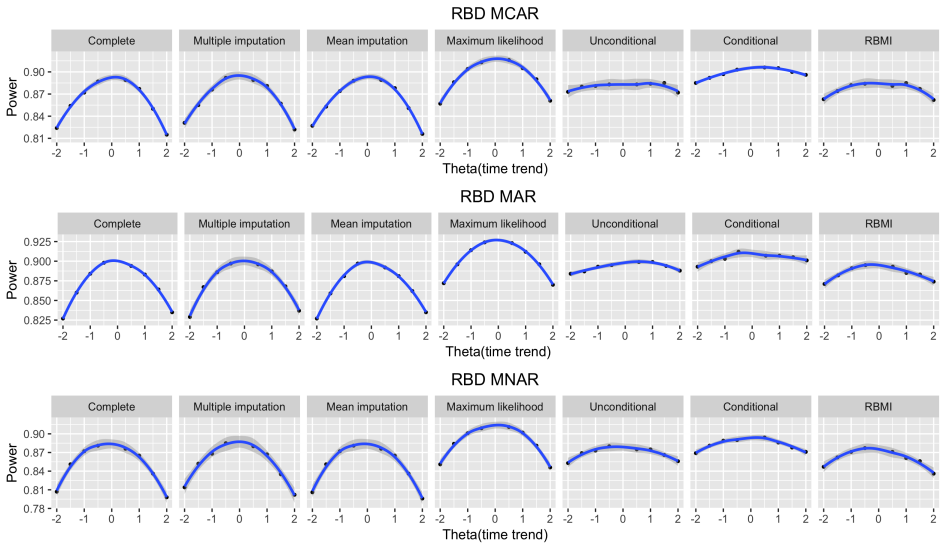

Figure A.3: *Time trend: RBD, power,* $p_{ms} = 0.1$*,* $n = 50$

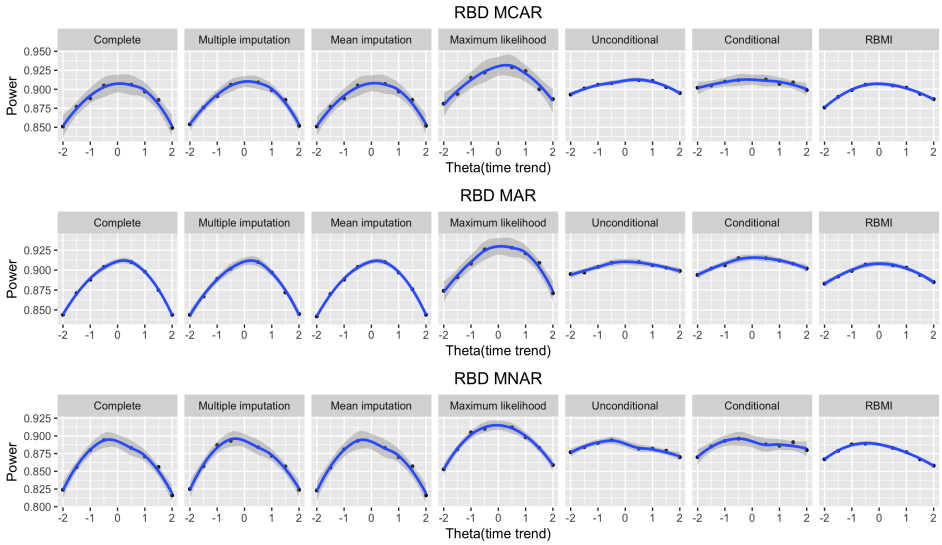

Figure A.4: *Time trend: RBD, power,* $p_{ms} = 0.1$*,* $n = 100$

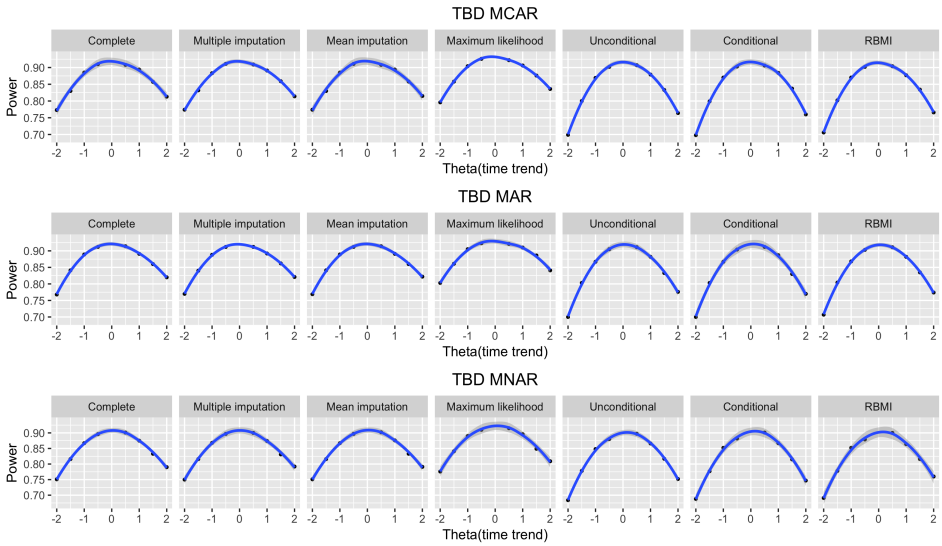

Figure A.5: *Time trend: TBD, power,* $p_{ms} = 0.05$*,* $n = 50$

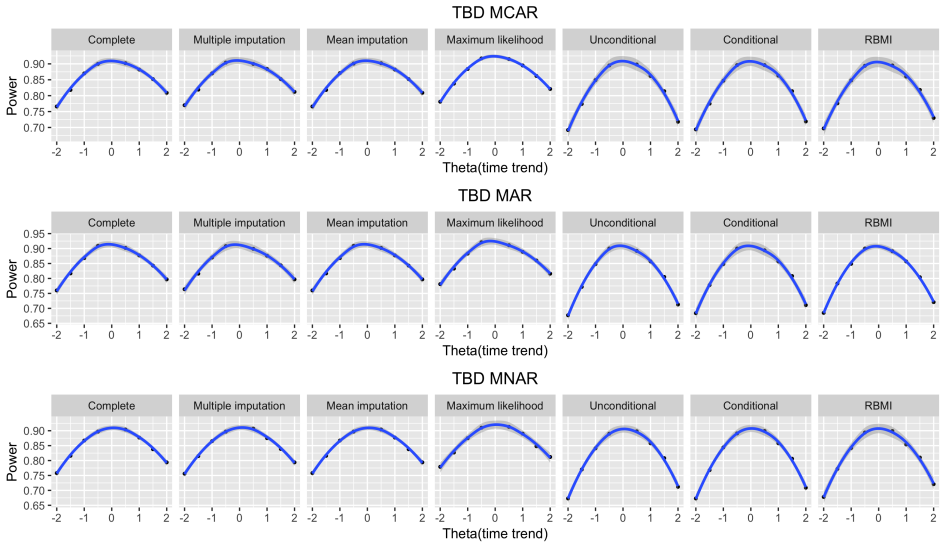

Figure A.6: *Time trend: TBD, power,* $p_{ms} = 0.05$*,* $n = 100$

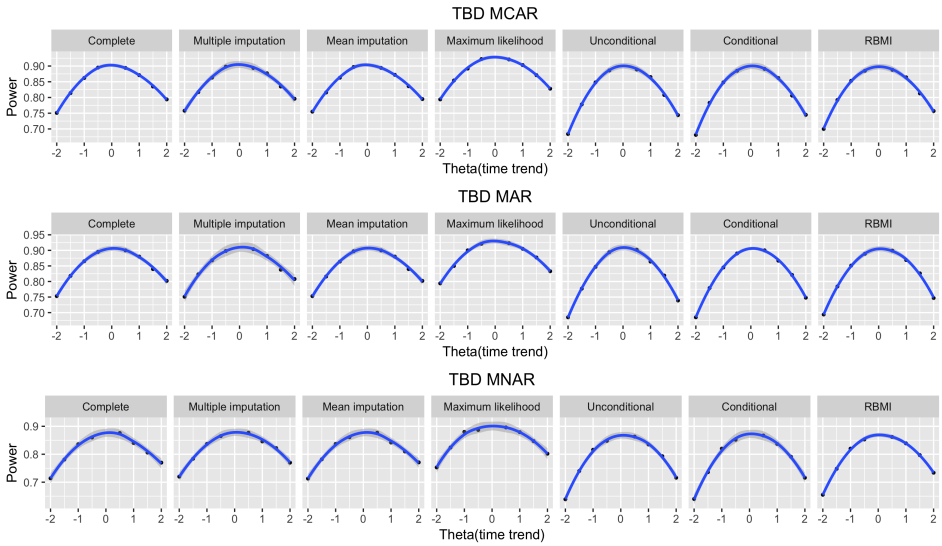

Figure A.7: *Time trend: TBD, power,* $p_{ms} = 0.1$*,* $n = 50$

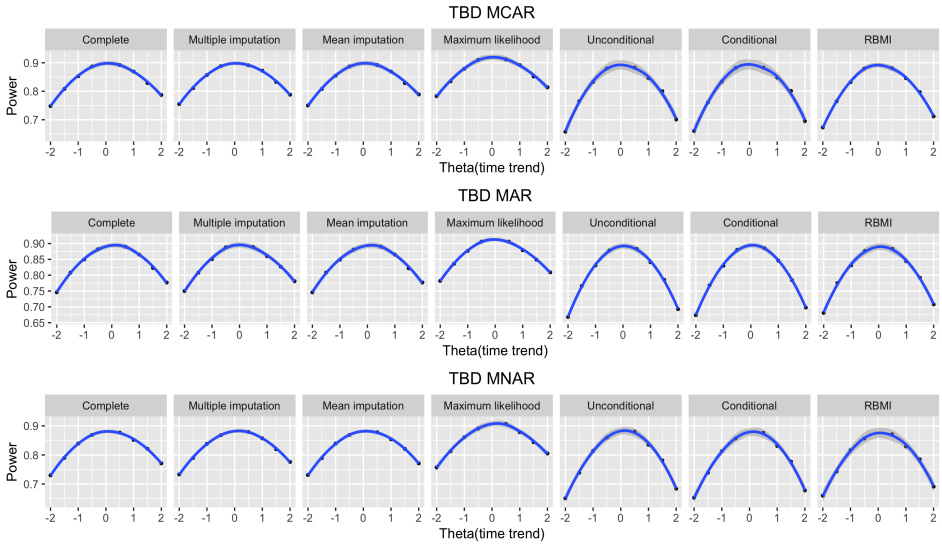

Figure A.8: *Time trend: TBD, power,* $p_{ms} = 0.1$, $n = 100$

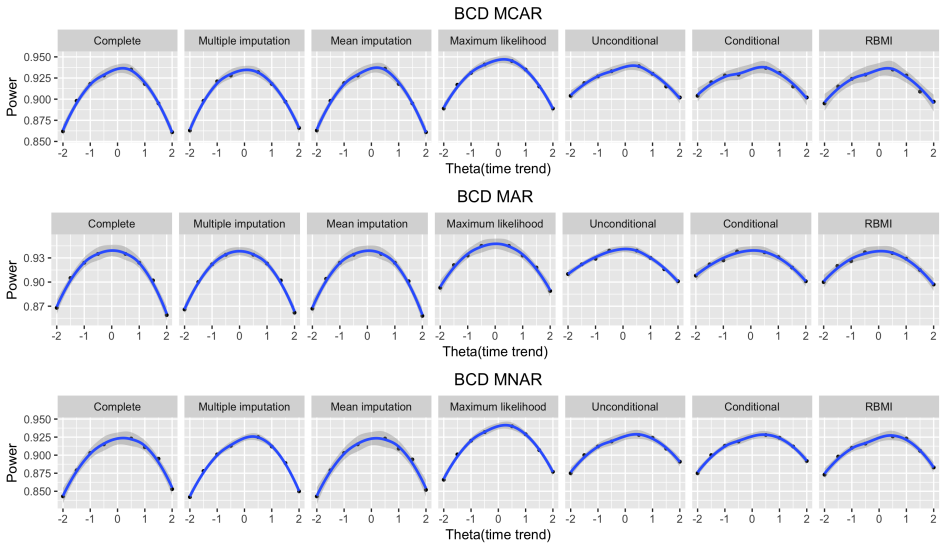

Figure A.9: *Time trend: BCD, power,* $p_{ms} = 0.05$, $n = 50$

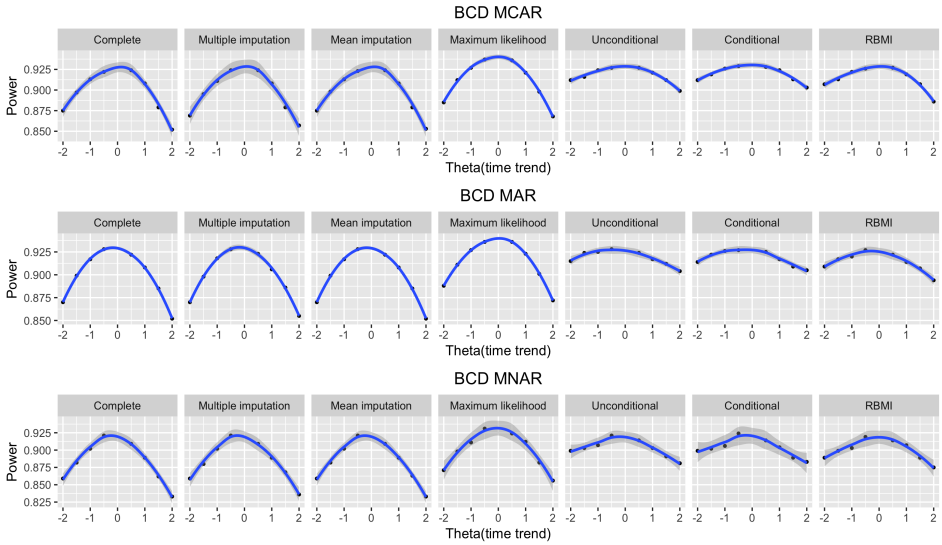

Figure A.10: *Time trend: BCD, power,* $p_{ms} = 0.05$*,* $n = 100$

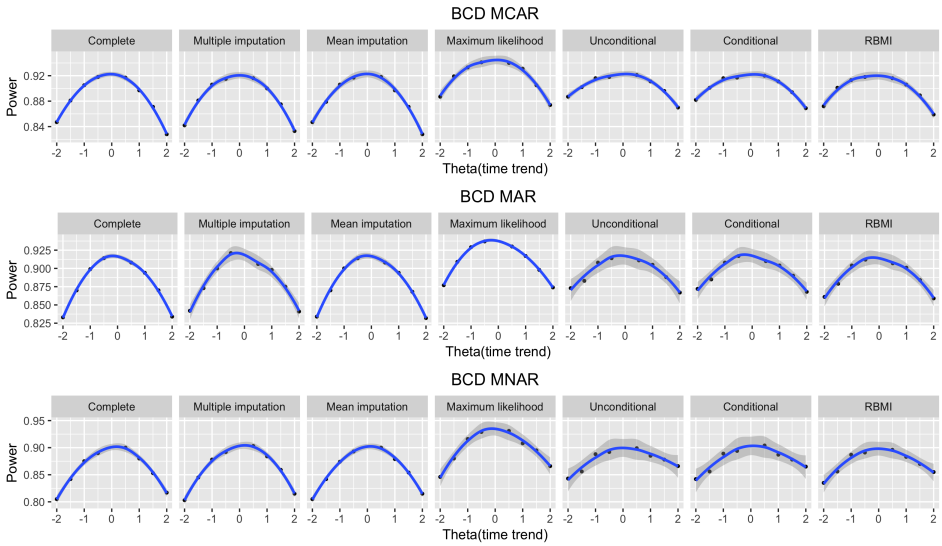

Figure A.11: *Time trend: BCD, power,* $p_{ms} = 0.1$*,* $n = 50$

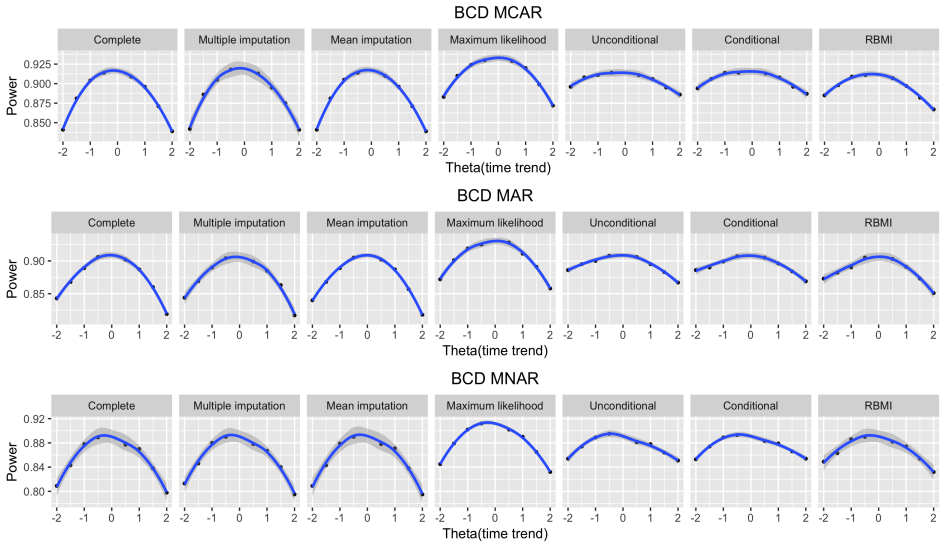

Figure A.12: *Time trend: BCD, power,* $p_{ms} = 0.1$*,* $n = 100$

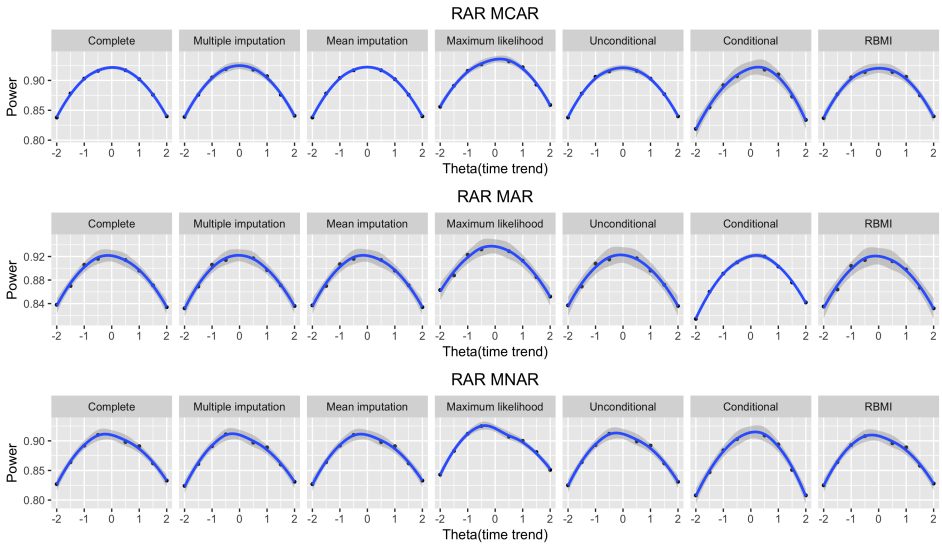

Figure A.13: *Time trend: RAR, power,* $p_{ms} = 0.05$*,* $n = 50$

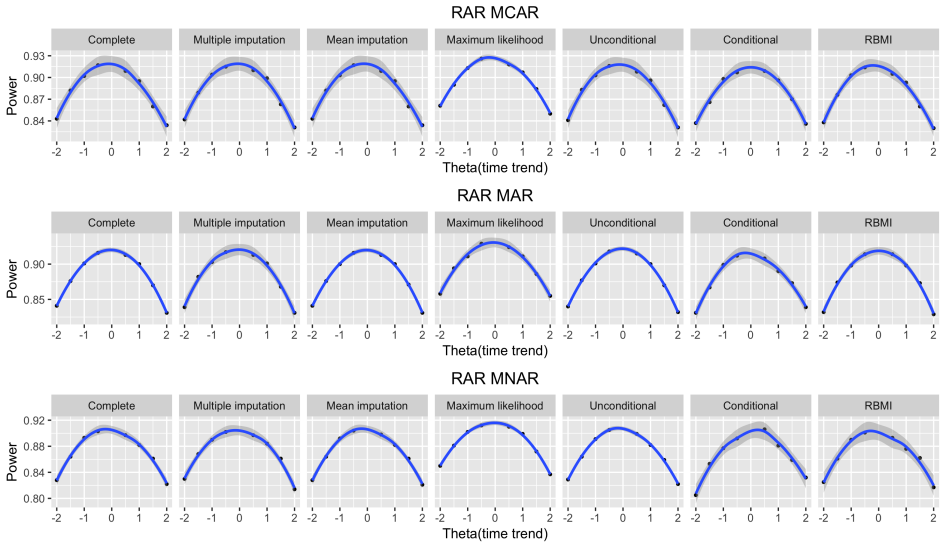

Figure A.14: *Time trend: RAR, power,* $p_{ms} = 0.05$*,* $n = 100$

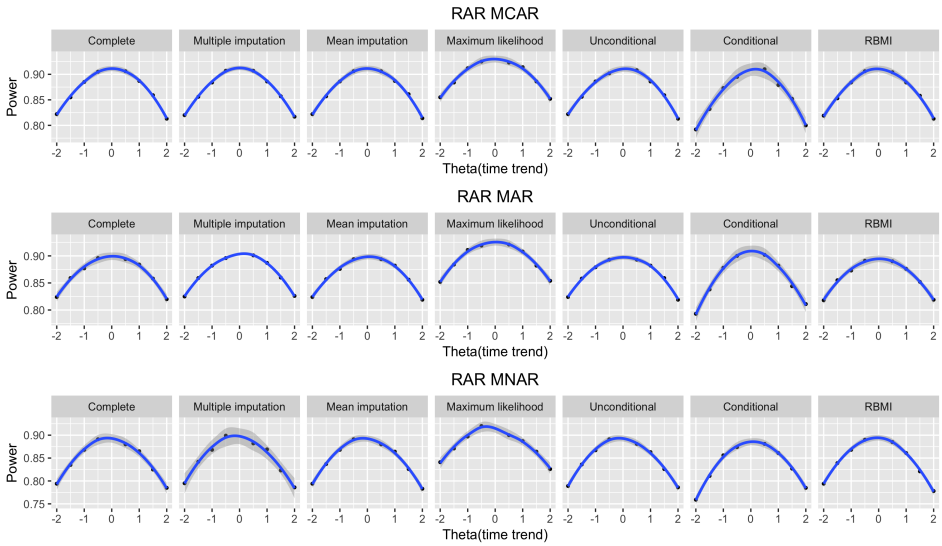

Figure A.15: *Time trend: RAR, power,* $p_{ms} = 0.1$*,* $n = 50$

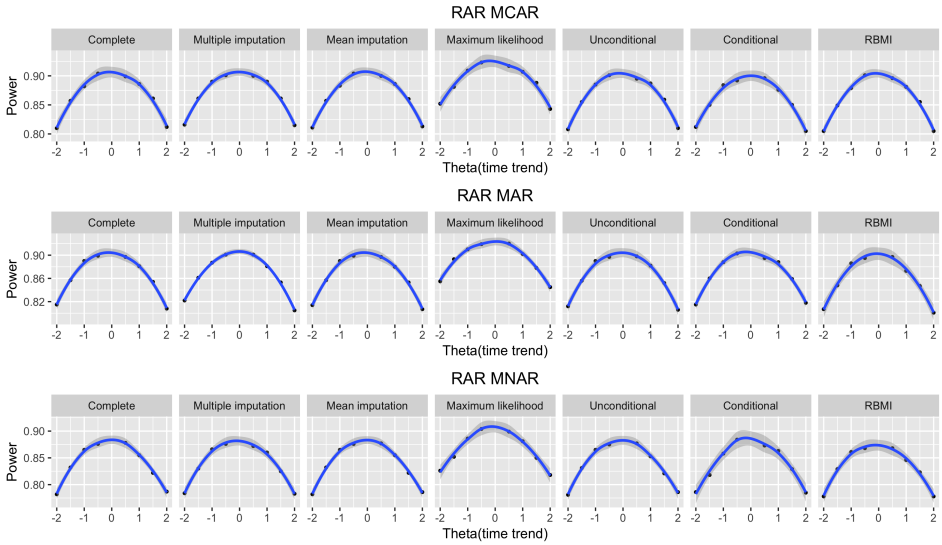

Figure A.16: *Time trend: RAR, power,* $p_{ms} = 0.1$*,* $n = 100$

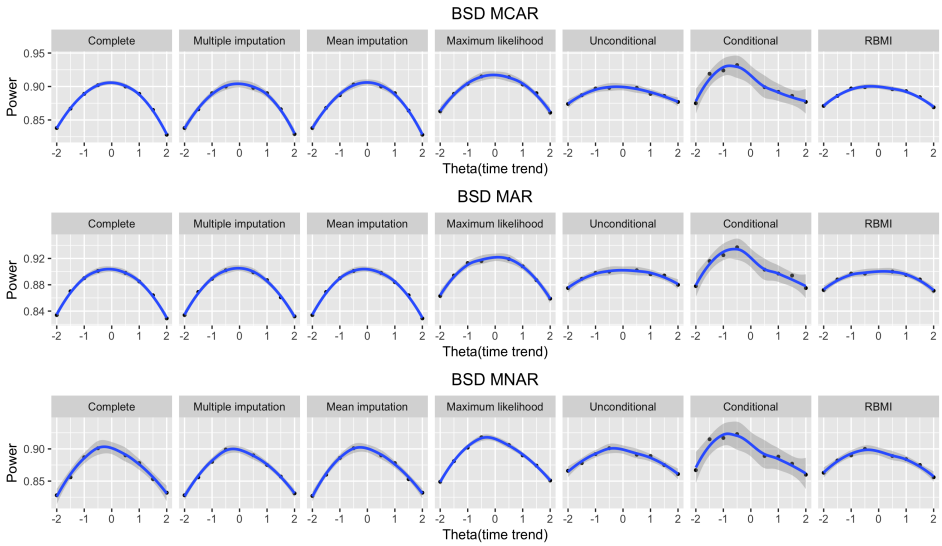

Figure A.17: *Time trend: BSD, power,* $p_{ms} = 0.05$*,* $n = 50$

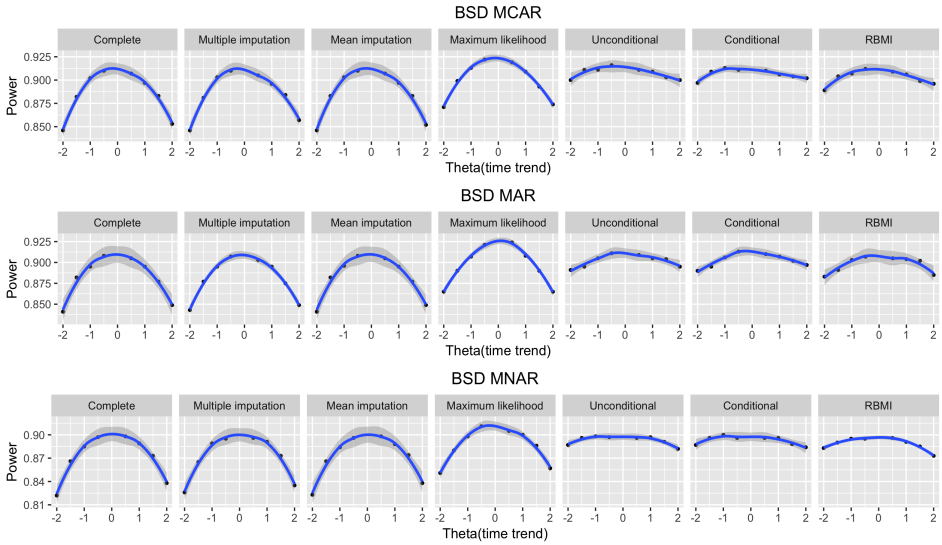

Figure A.18: *Time trend: BSD, power,* $p_{ms} = 0.05$*,* $n = 100$

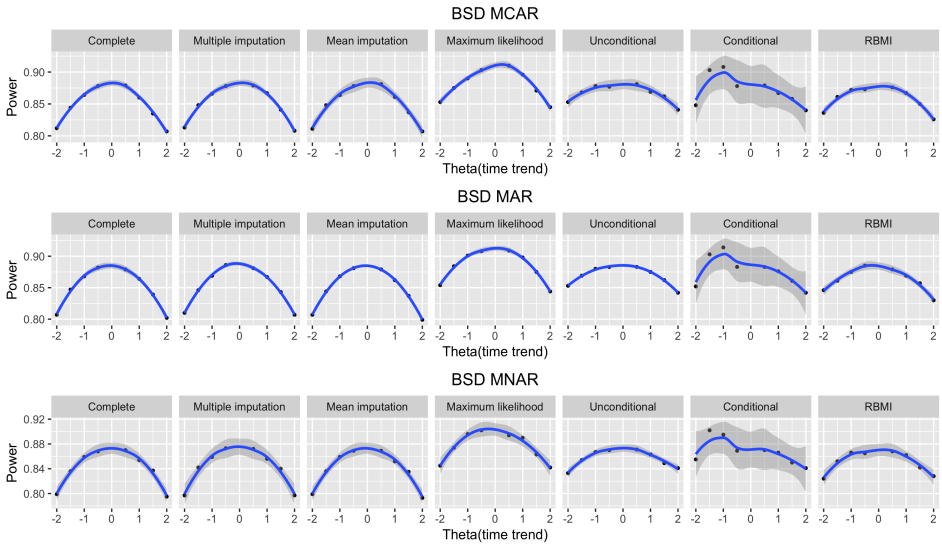

Figure A.19: *Time trend: BSD, power,* $p_{ms} = 0.1$*,* $n = 50$

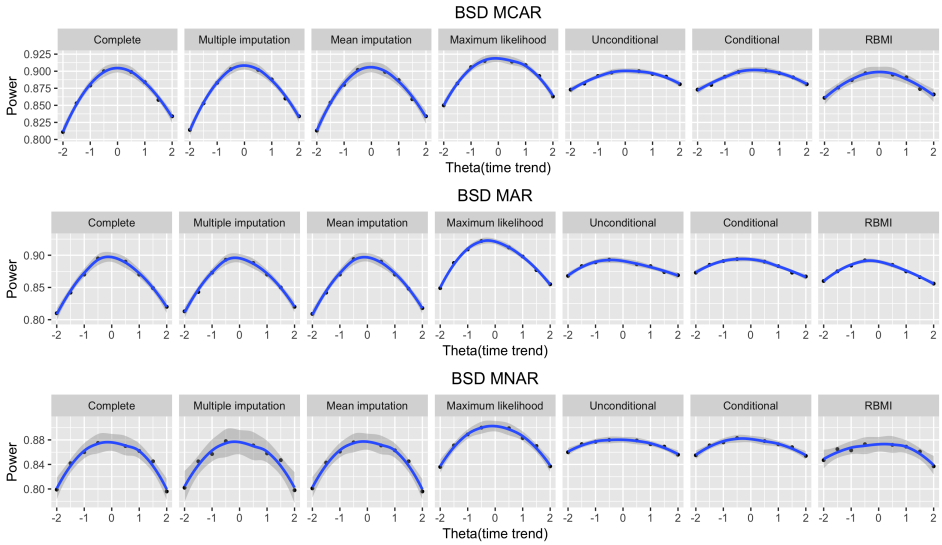

Figure A.20: *Time trend: BSD, power,* $p_{ms} = 0.1$*,* $n = 100$

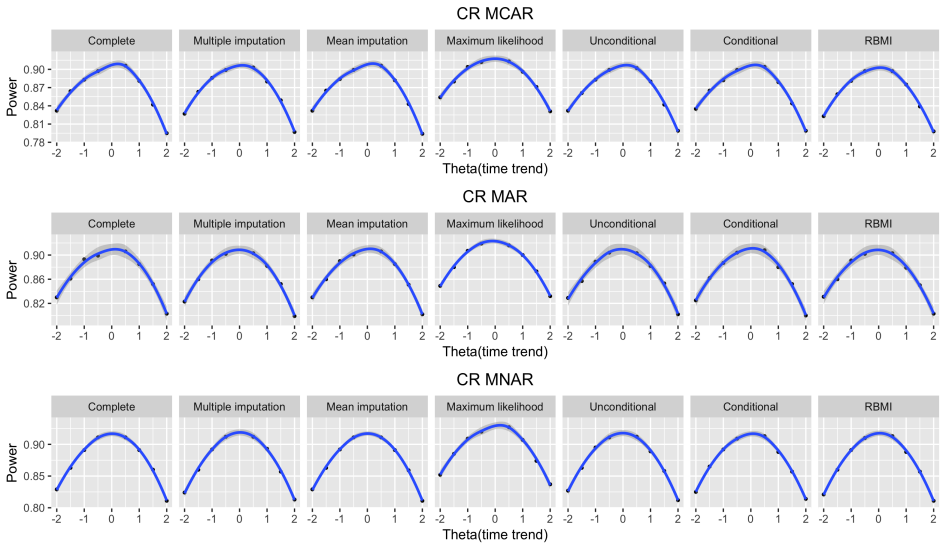

Figure A.21: *Time trend: CR, power,* $p_{ms} = 0.05$*,* $n = 50$

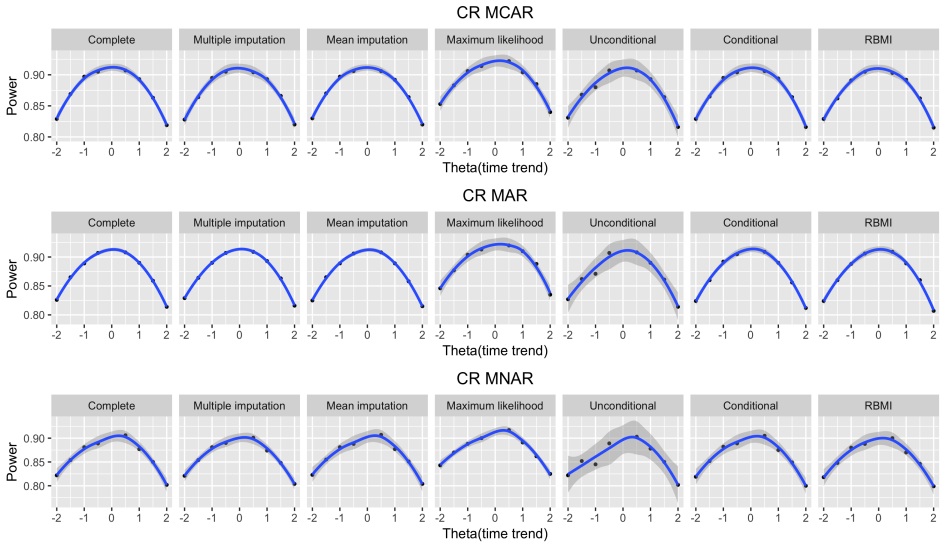

Figure A.22: *Time trend: CR, power,* $p_{ms} = 0.05$*,* $n = 100$

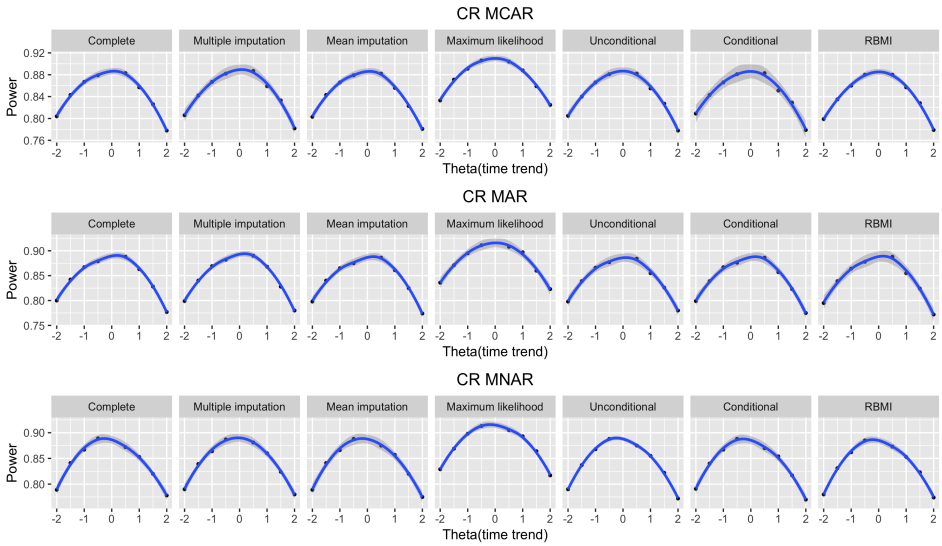

Figure A.23: *Time trend: CR, power,* $p_{ms} = 0.1$*,* $n = 50$

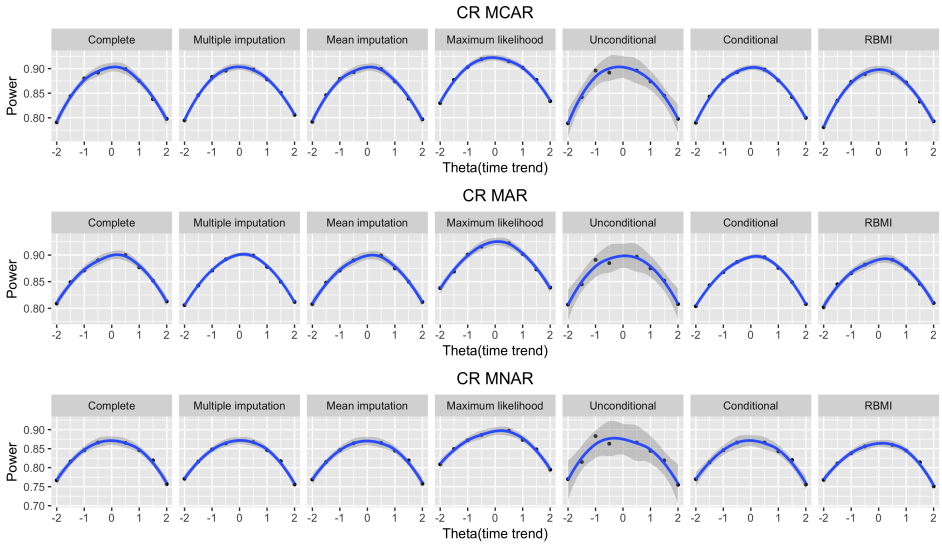

Figure A.24: *Time trend: CR, power,* $p_{ms} = 0.1$*,* $n = 100$

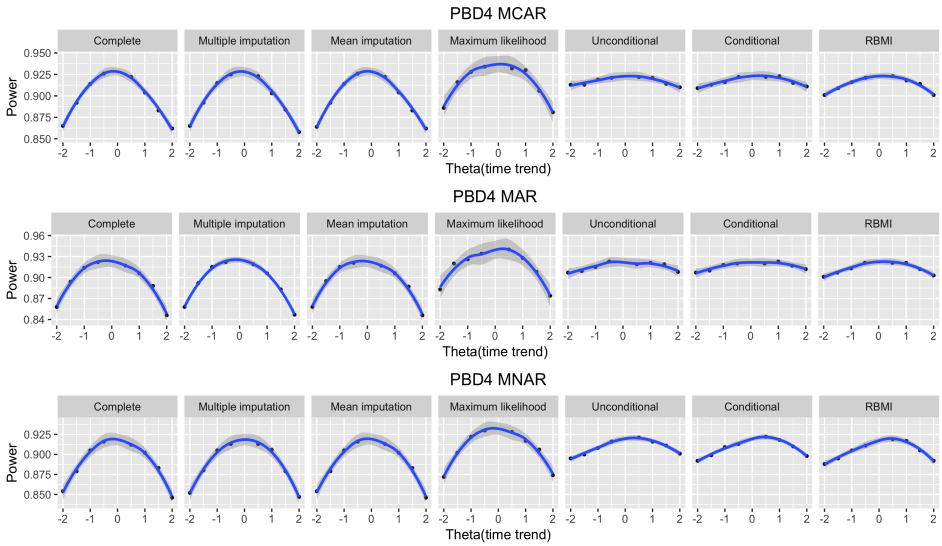

Figure A.25: *Time trend: PBD (blocksize = 4), power,* $p_{ms} = 0.05$*,* $n = 50$

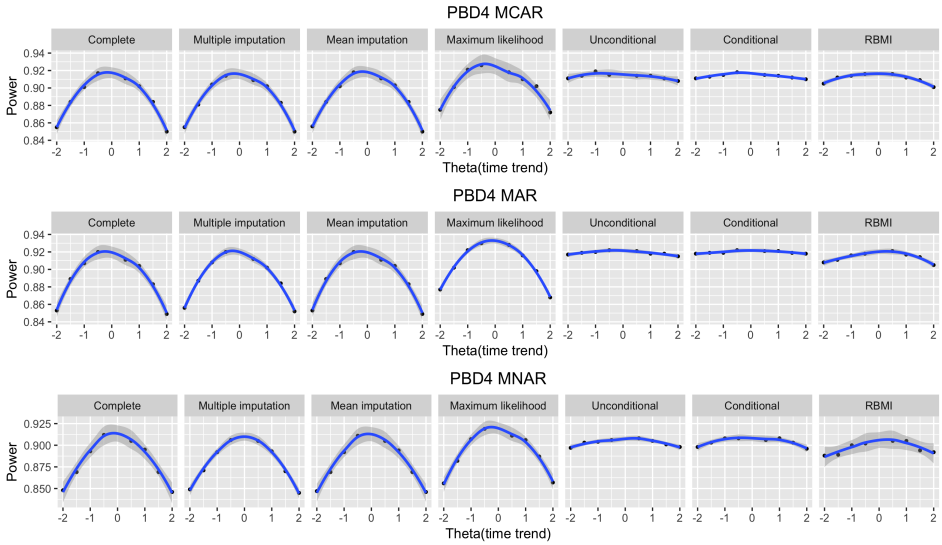

Figure A.26: *Time trend: PBD (blocksize = 4), power,* $p_{ms} = 0.05$, $n = 100$

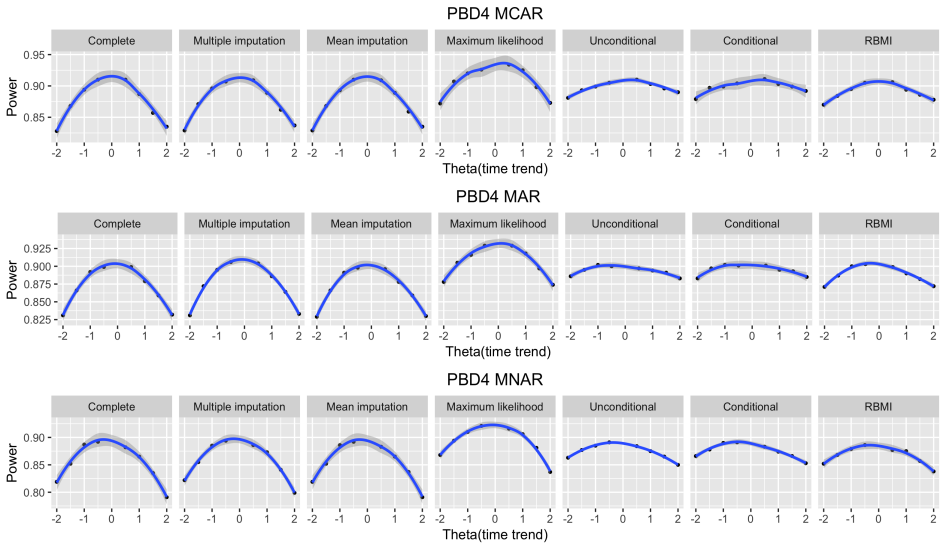

Figure A.27: *Time trend: PBD (blocksize = 4), power,* $p_{ms} = 0.1$, $n = 50$

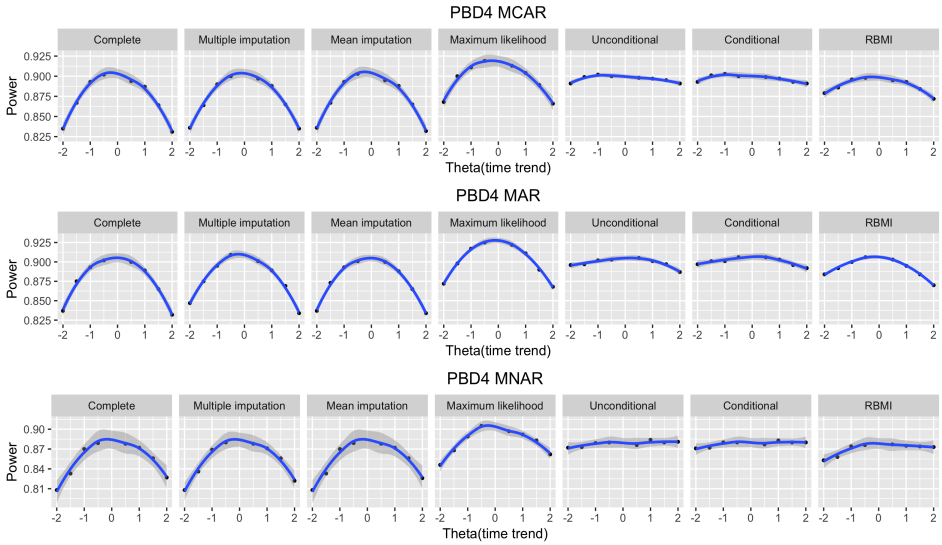

Figure A.28: *Time trend: PBD (blocksize = 4), power,* $p_{ms} = 0.1$*,* $n = 100$

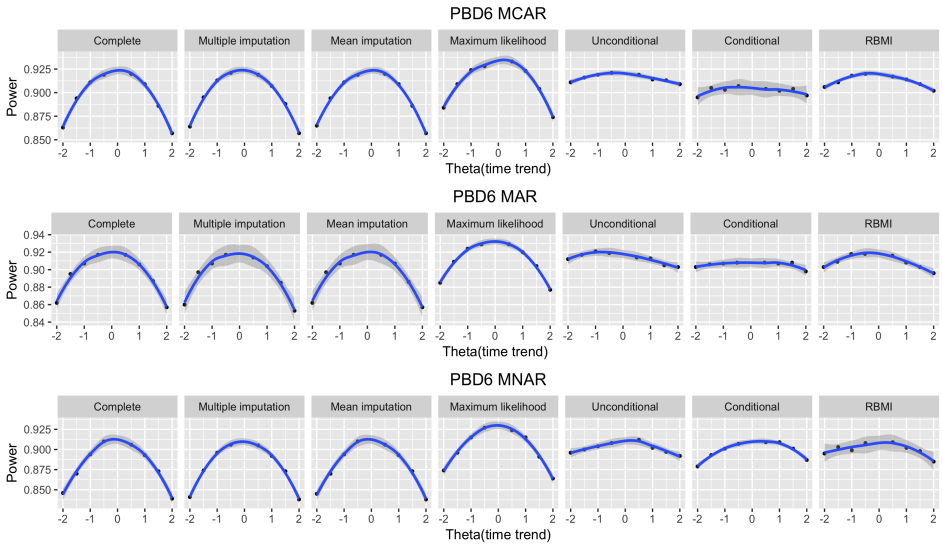

Figure A.29: *Time trend: PBD (blocksize = 6), power,* $p_{ms} = 0.05$*,* $n = 50$

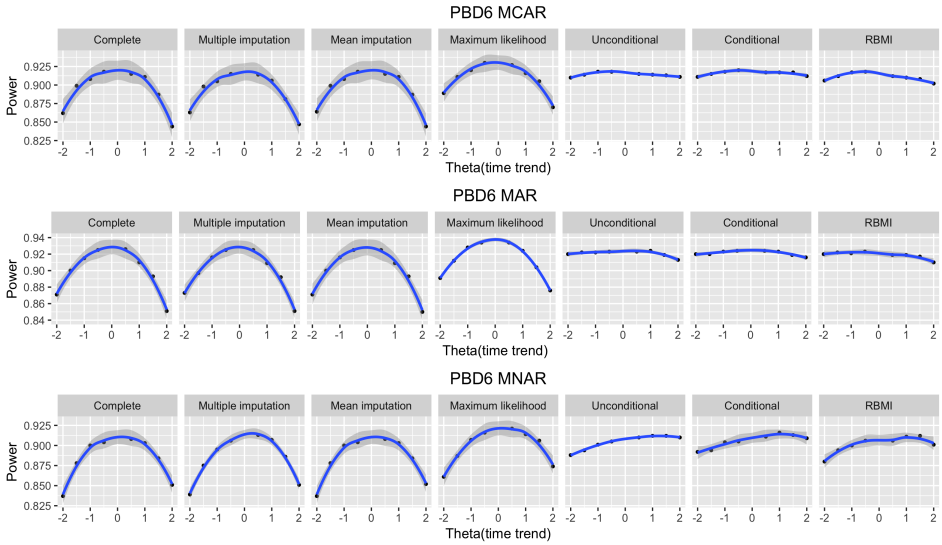

Figure A.30: *Time trend: PBD (blocksize = 6), power,* $p_{ms} = 0.05$, $n = 100$

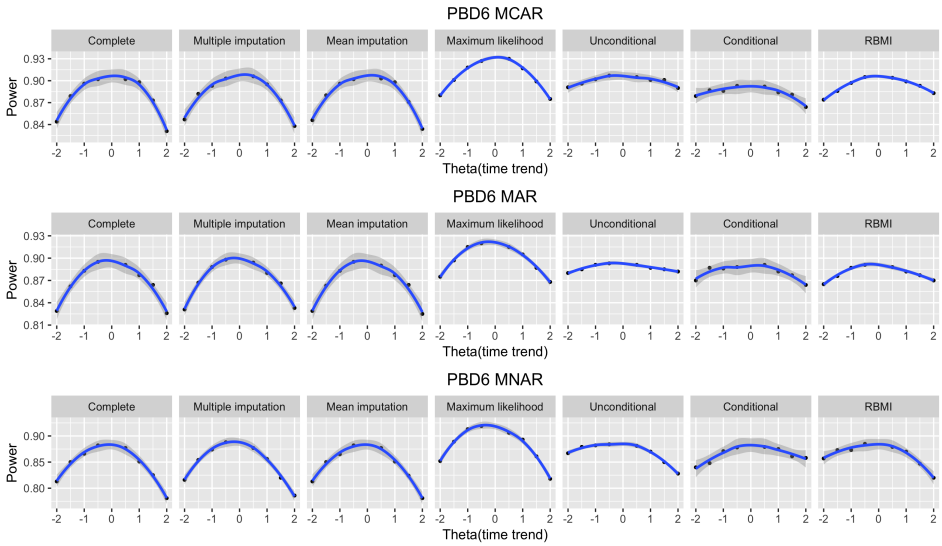

Figure A.31: *Time trend: PBD (blocksize = 6), power,* $p_{ms} = 0.1$, $n = 50$

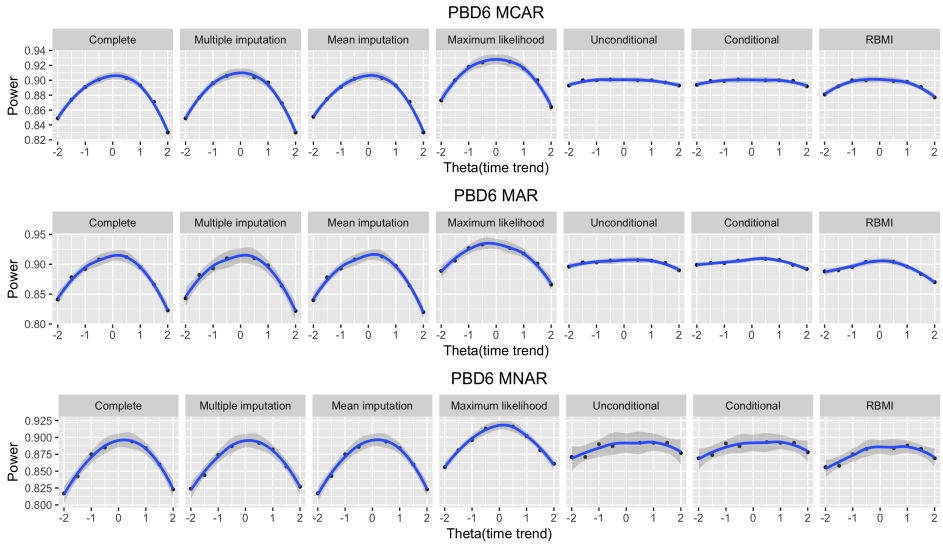

Figure A.32: *Time trend: PBD (blocksize = 6), power,* $p_{ms} = 0.1$, $n = 100$

A.3.2 Type I Error Rates

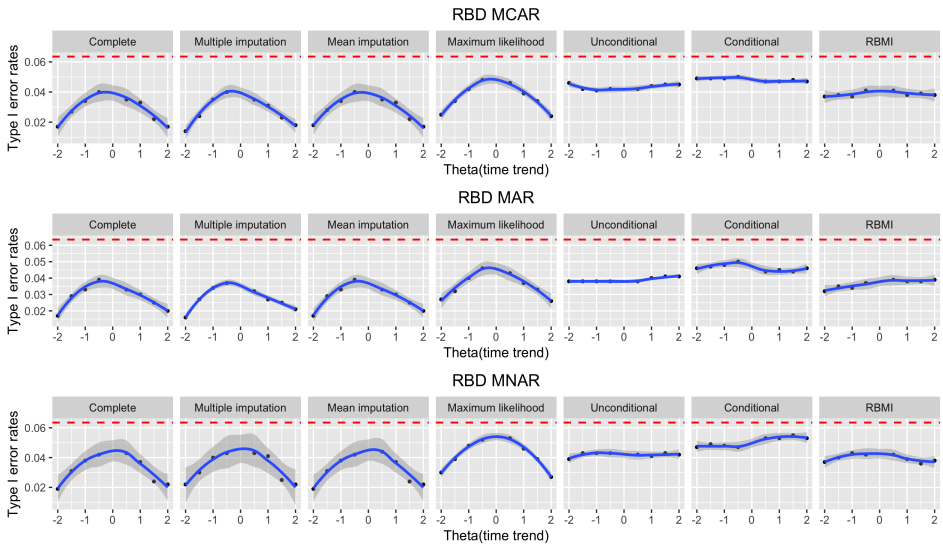

Figure A.33: *Time trend: RBD, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

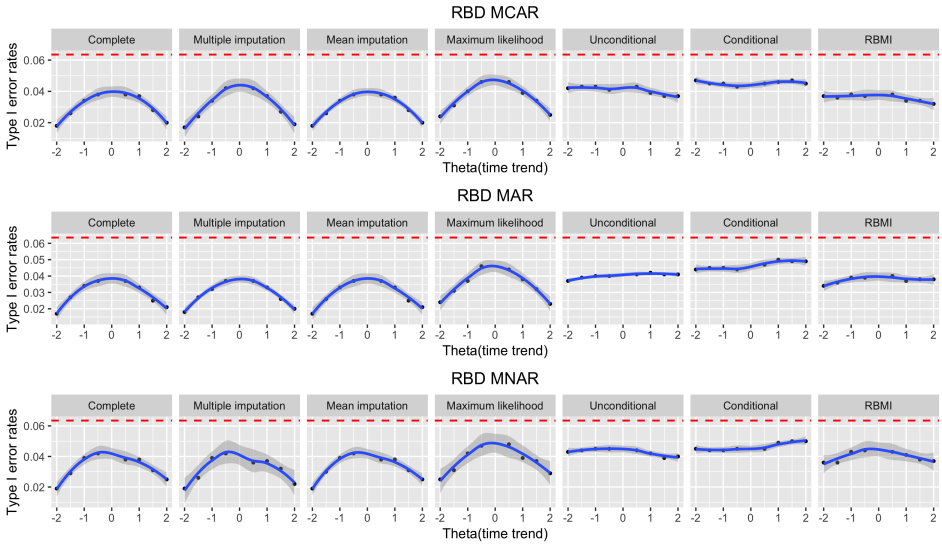

Figure A.34: *Time trend: RBD, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

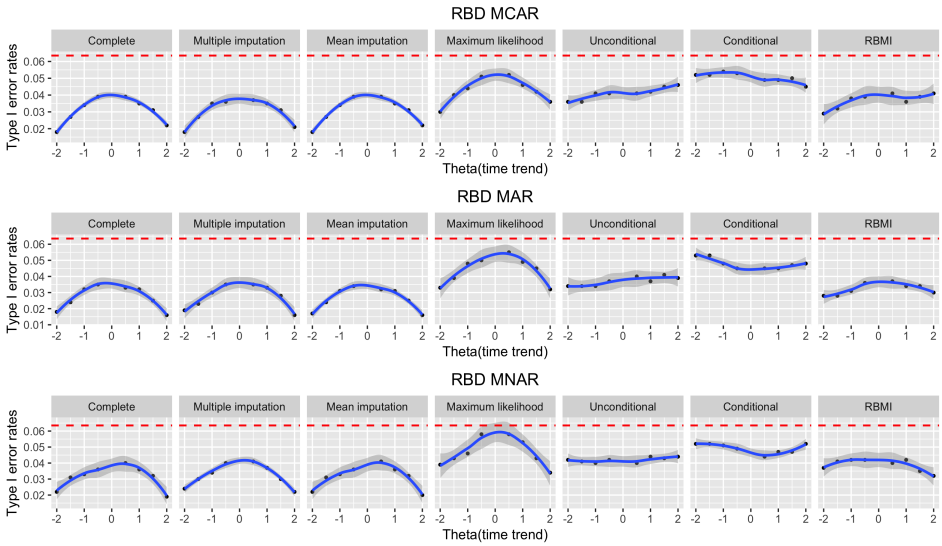

Figure A.35: *Time trend: RBD, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

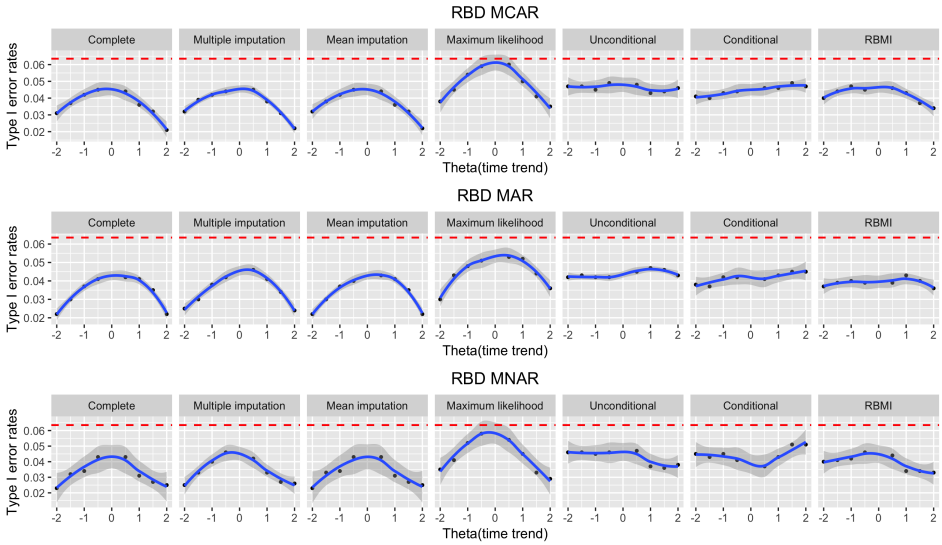

Figure A.36: *Time trend: RBD, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

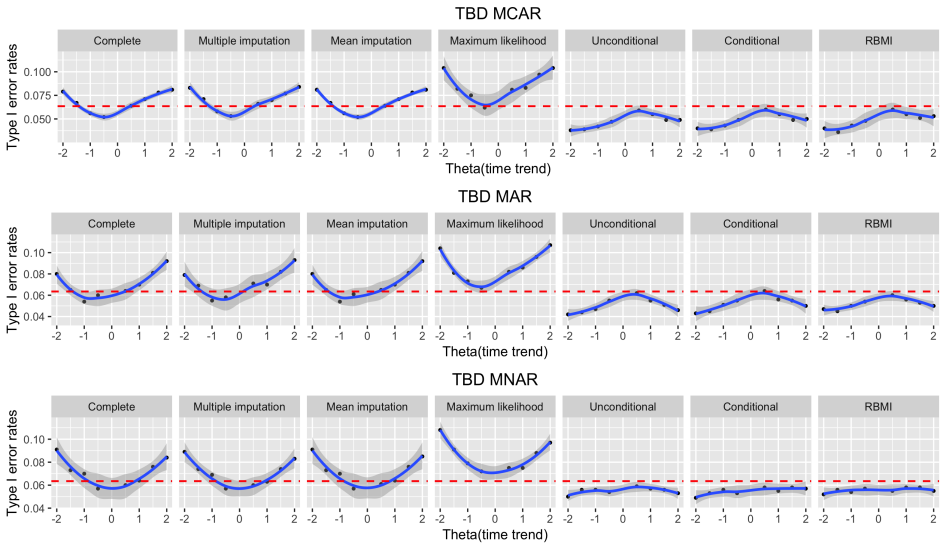

Figure A.37: *Time trend: TBD, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

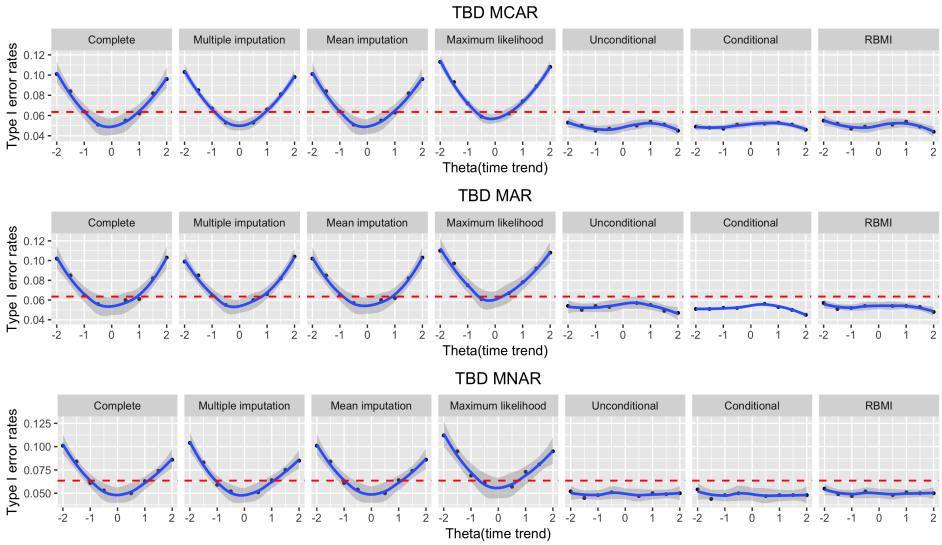

Figure A.38: *Time trend: TBD, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

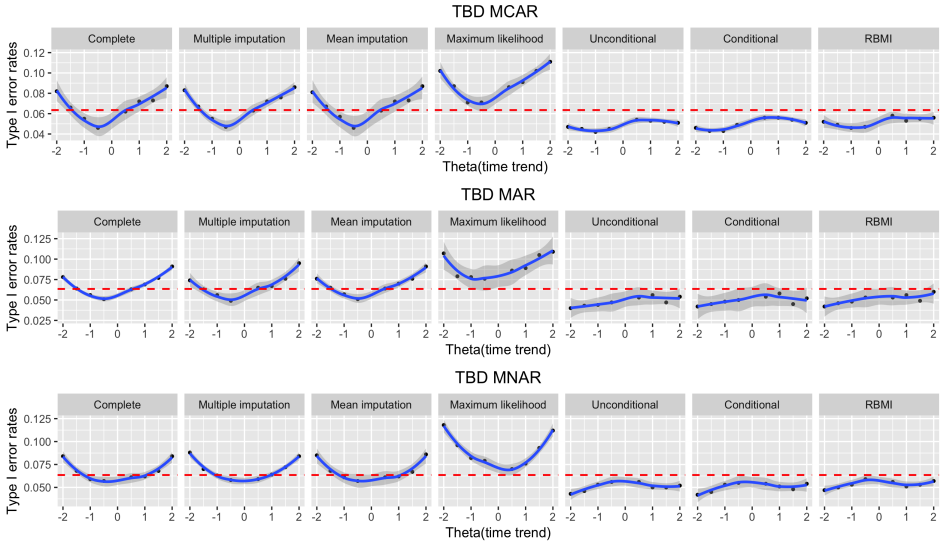

Figure A.39: *Time trend: TBD, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

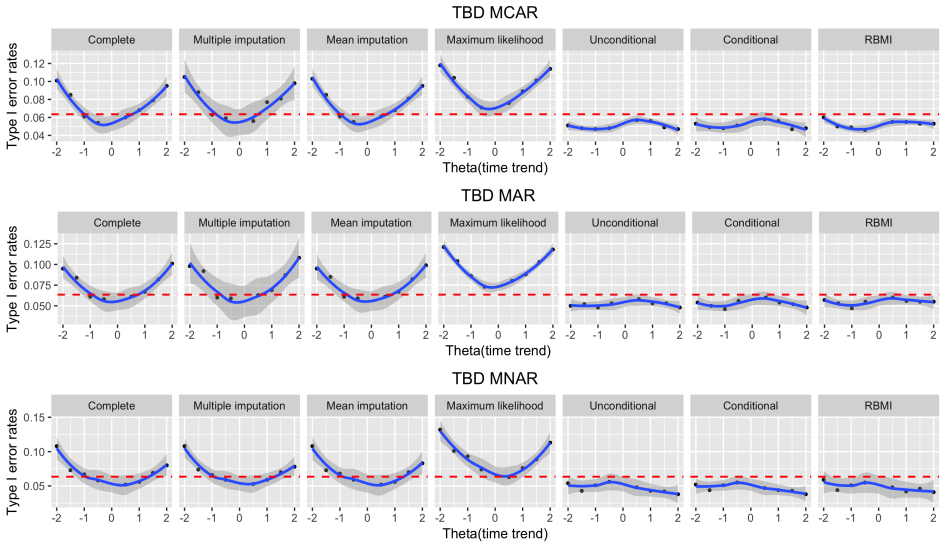

Figure A.40: *Time trend: TBD, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

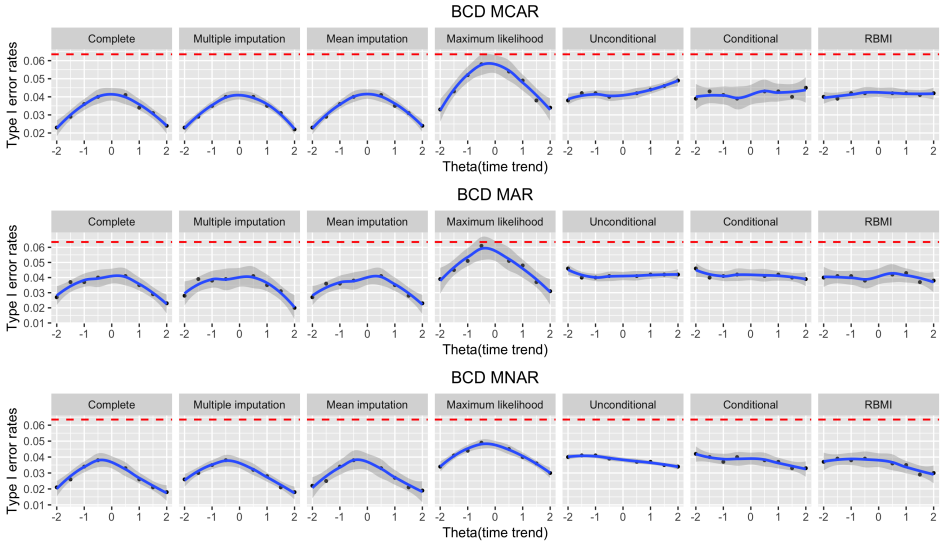

Figure A.41: *Time trend: BCD, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

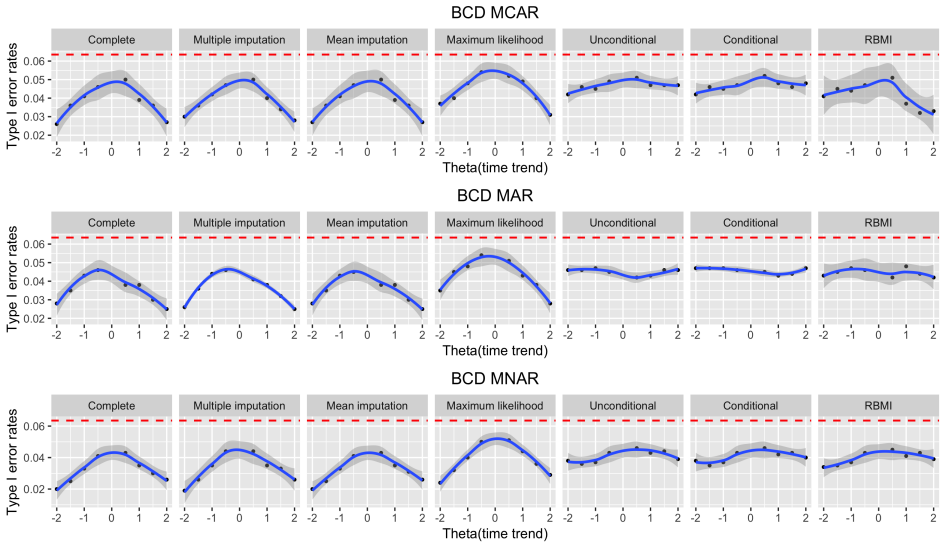

Figure A.42: *Time trend: BCD, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

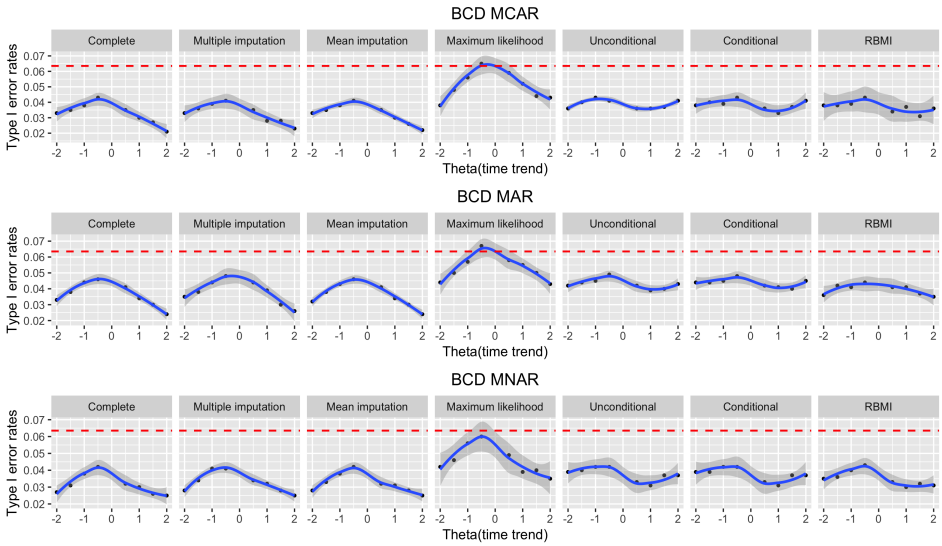

Figure A.43: *Time trend: BCD, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

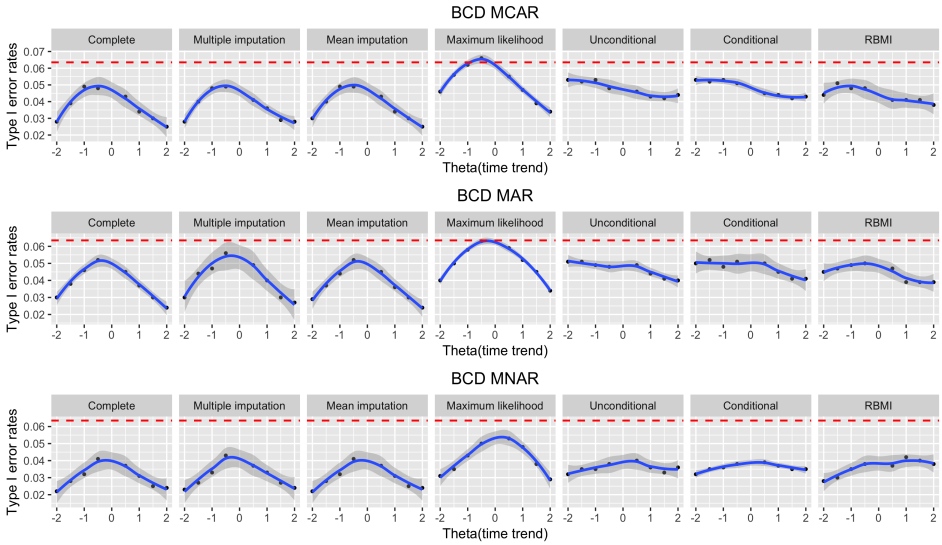

Figure A.44: *Time trend: BCD, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

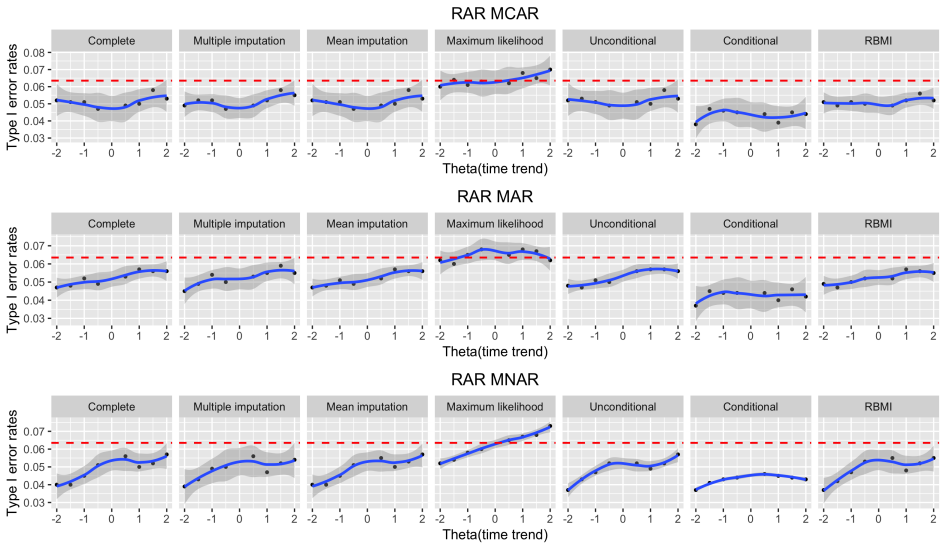

Figure A.45: *Time trend: RAR, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

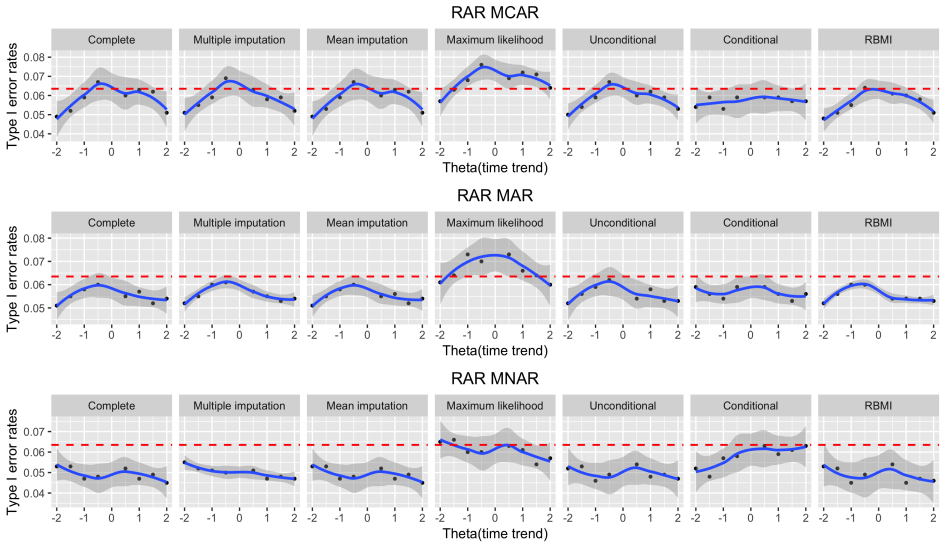

Figure A.46: *Time trend: RAR, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

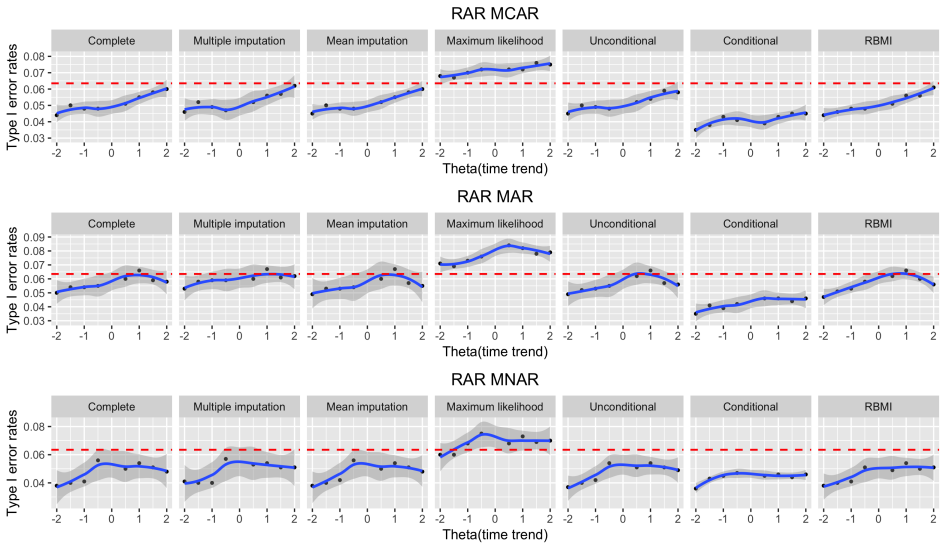

Figure A.47: *Time trend: RAR, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

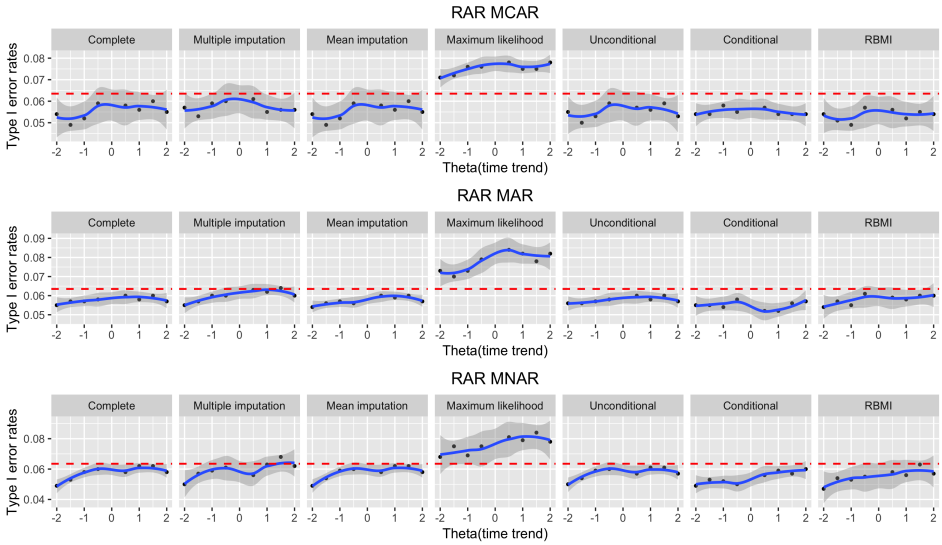

Figure A.48: *Time trend: RAR, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

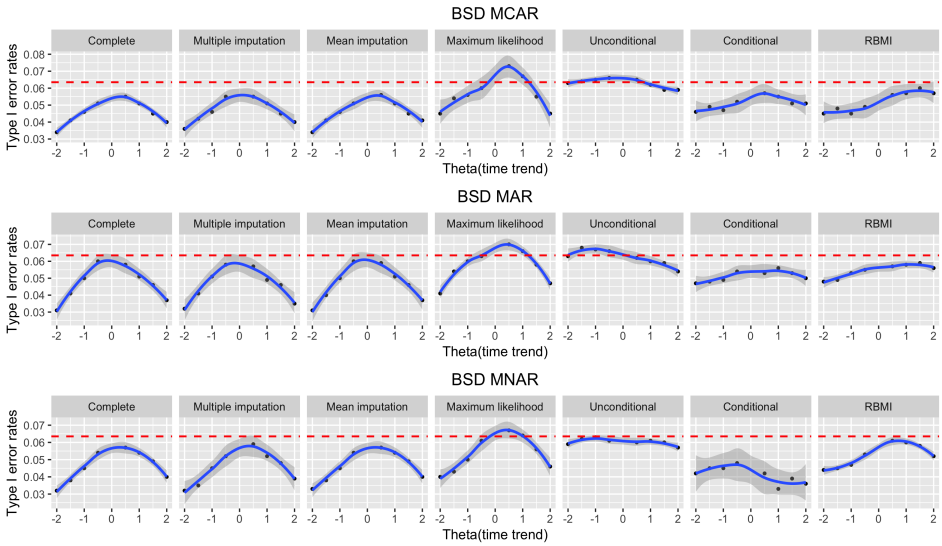

Figure A.49: *Time trend: BSD, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

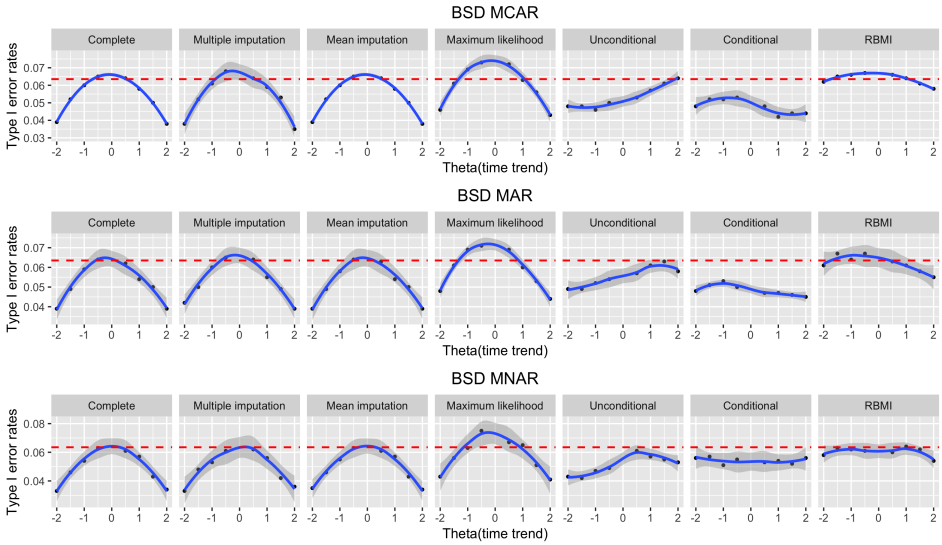

Figure A.50: *Time trend: BSD, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

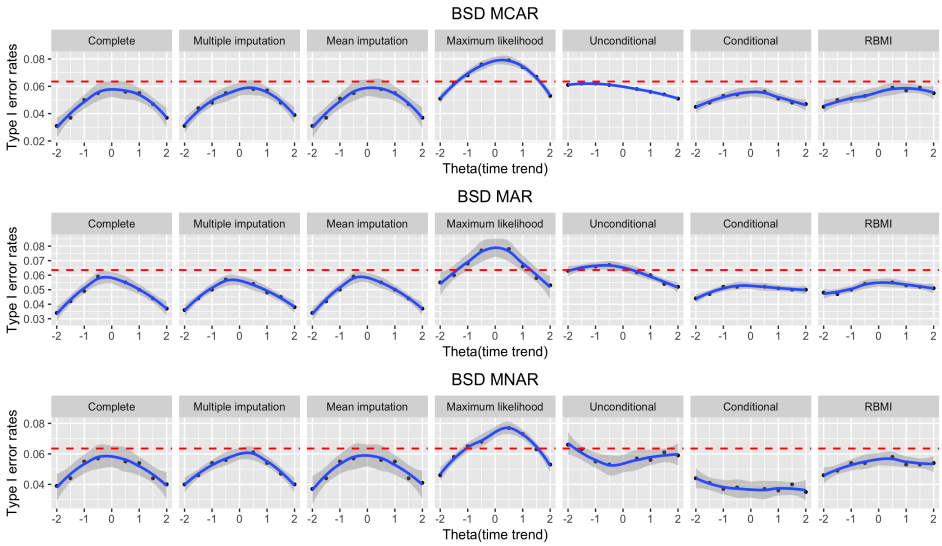

Figure A.51: *Time trend: BSD, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

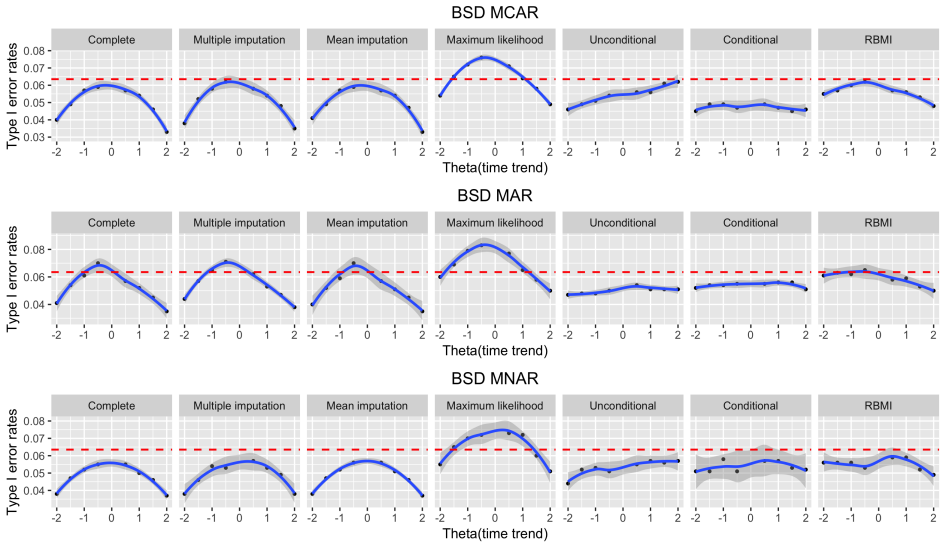

Figure A.52: *Time trend: BSD, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

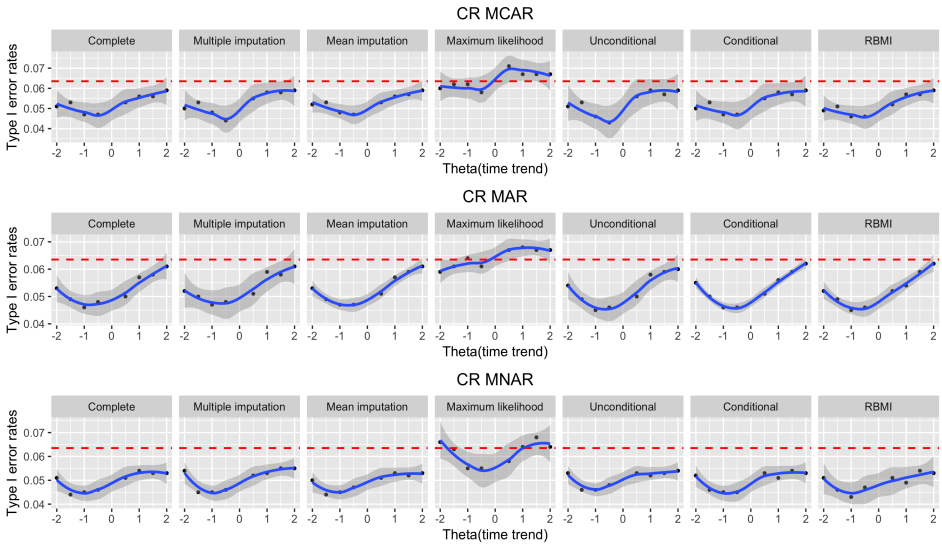

Figure A.53: *Time trend: CR, type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

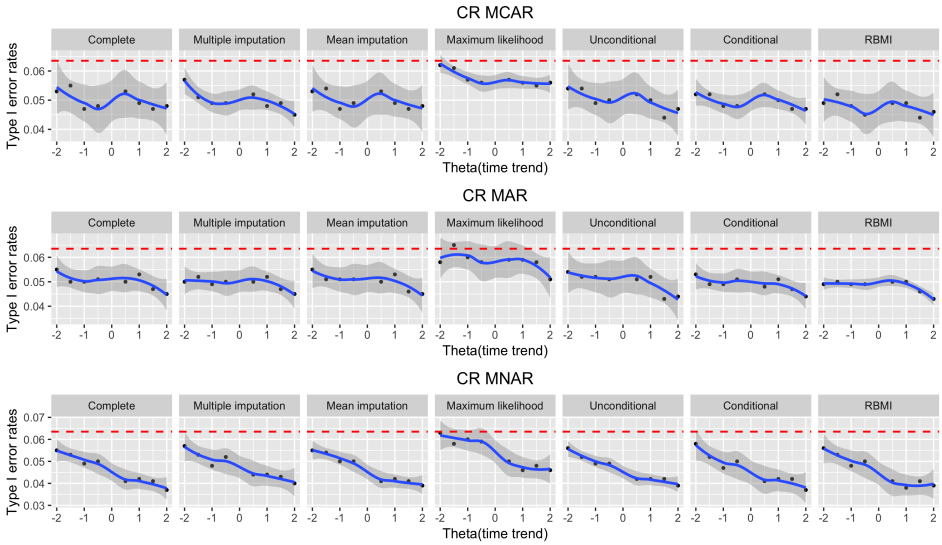

Figure A.54: *Time trend: CR, type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

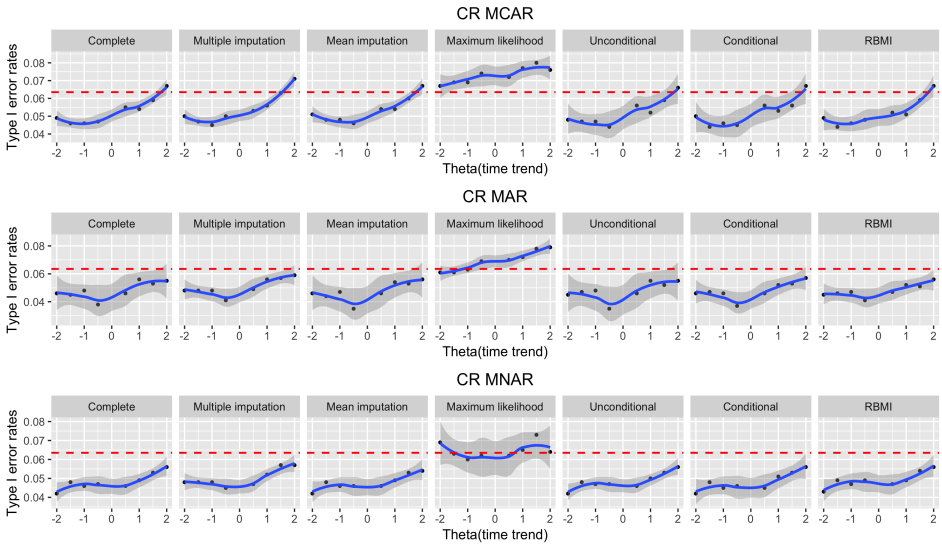

Figure A.55: *Time trend: CR, type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

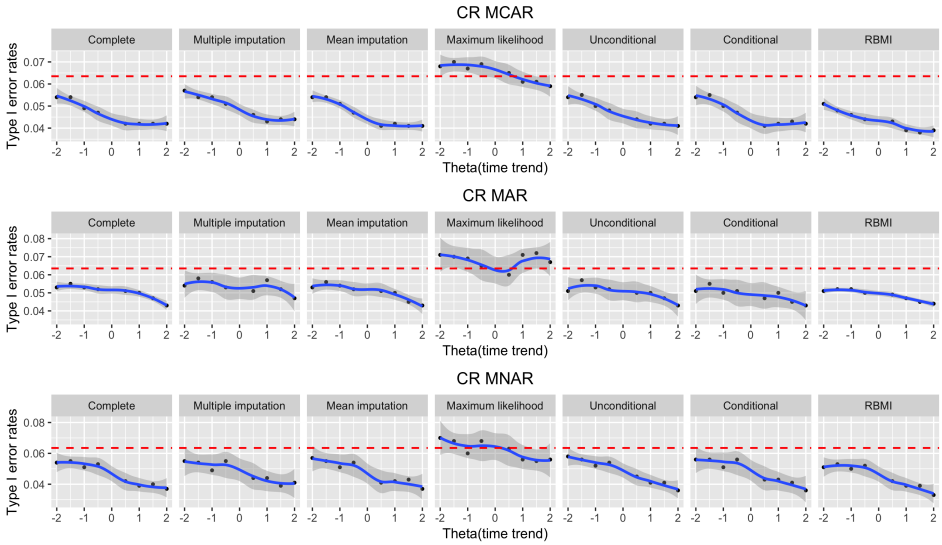

Figure A.56: *Time trend: CR, type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

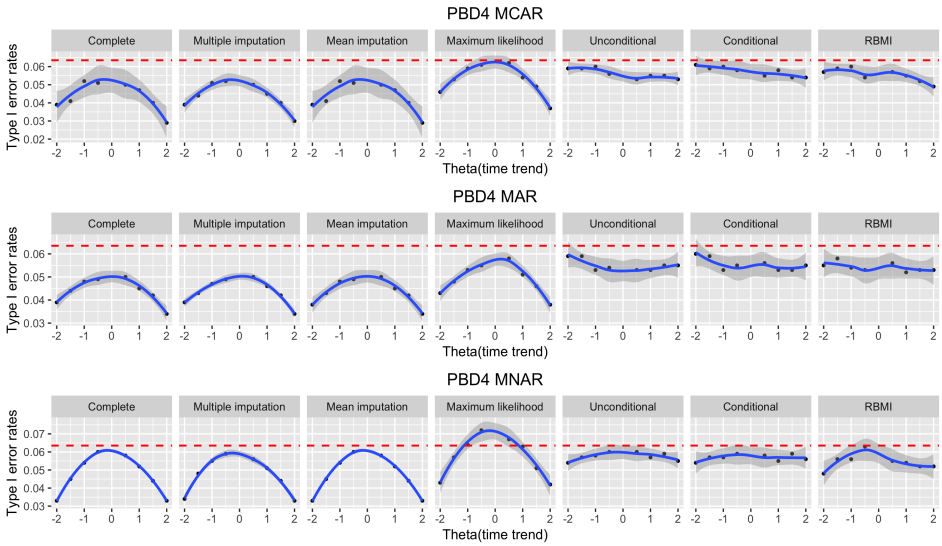

Figure A.57: *Time trend: PBD (blocksize = 4), type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

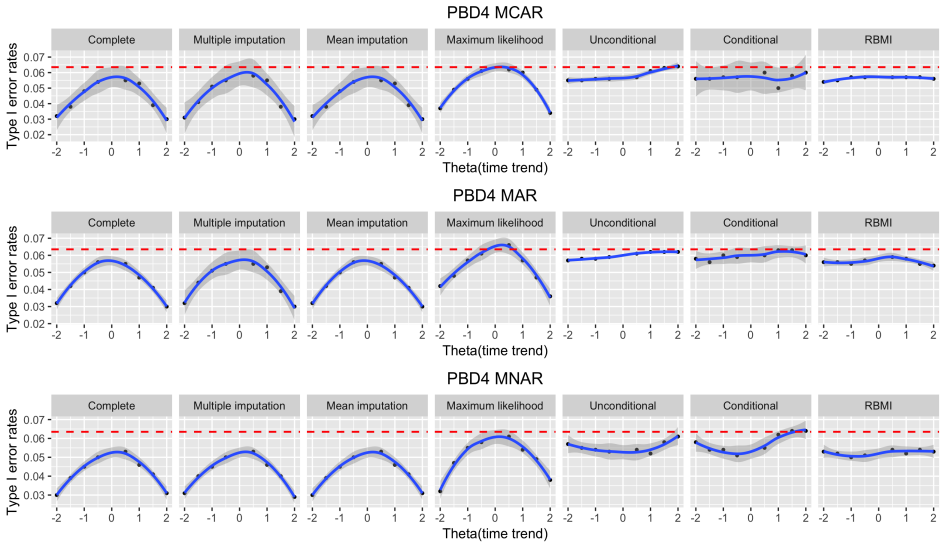

Figure A.58: *Time trend: PBD (blocksize = 4), type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

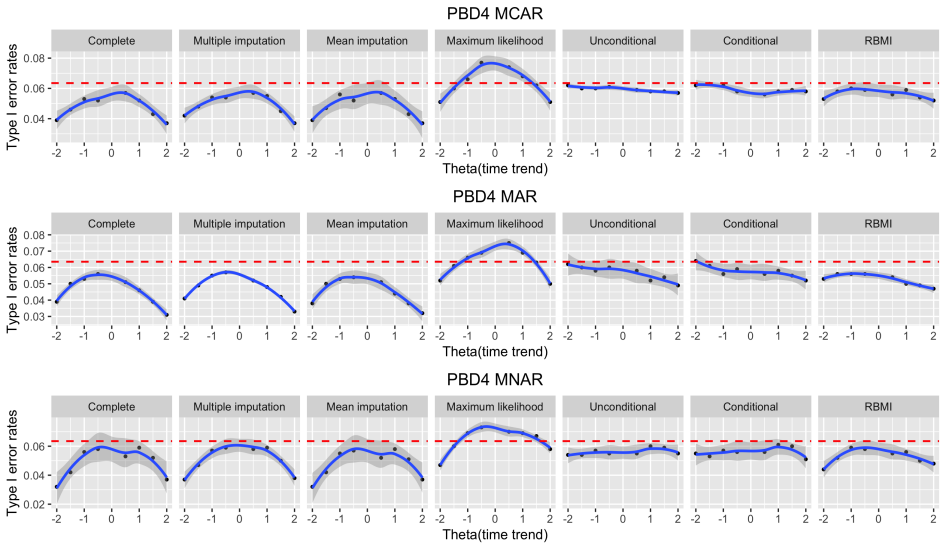

Figure A.59: *Time trend: PBD (blocksize = 4), type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

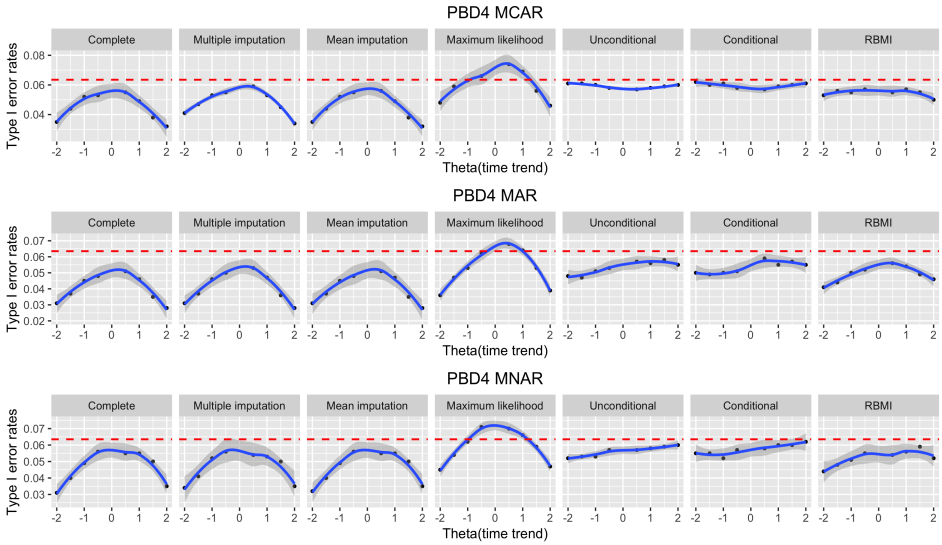

Figure A.60: *Time trend: PBD (blocksize = 4), type I error rates,* $p_{ms} = 0.1$*,* $n = 100$

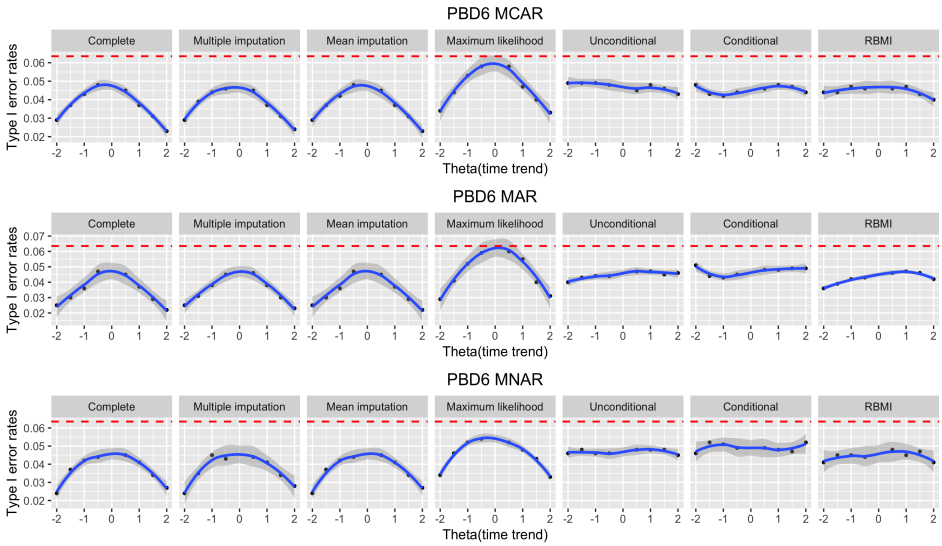

Figure A.61: *Time trend: PBD (blocksize = 6), type I error rates,* $p_{ms} = 0.05$*,* $n = 50$

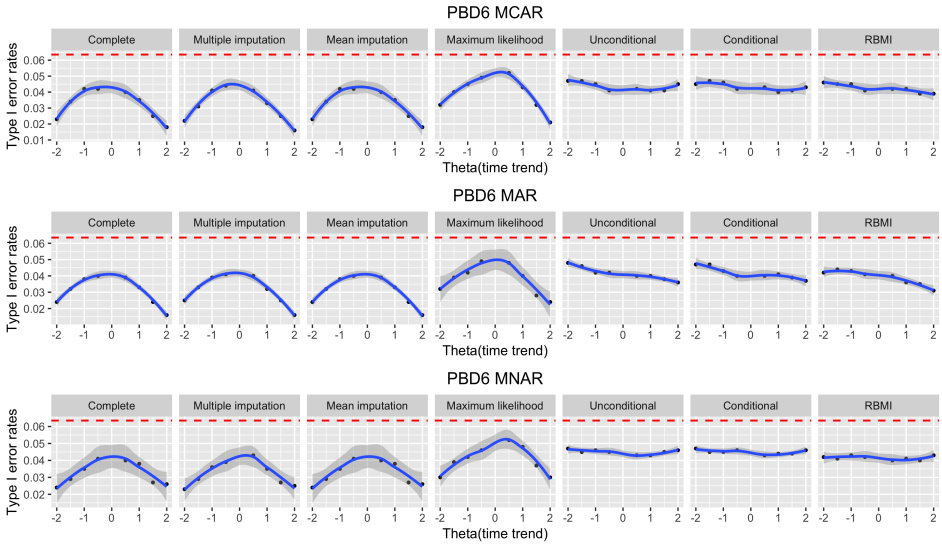

Figure A.62: *Time trend: PBD (blocksize = 6), type I error rates,* $p_{ms} = 0.05$*,* $n = 100$

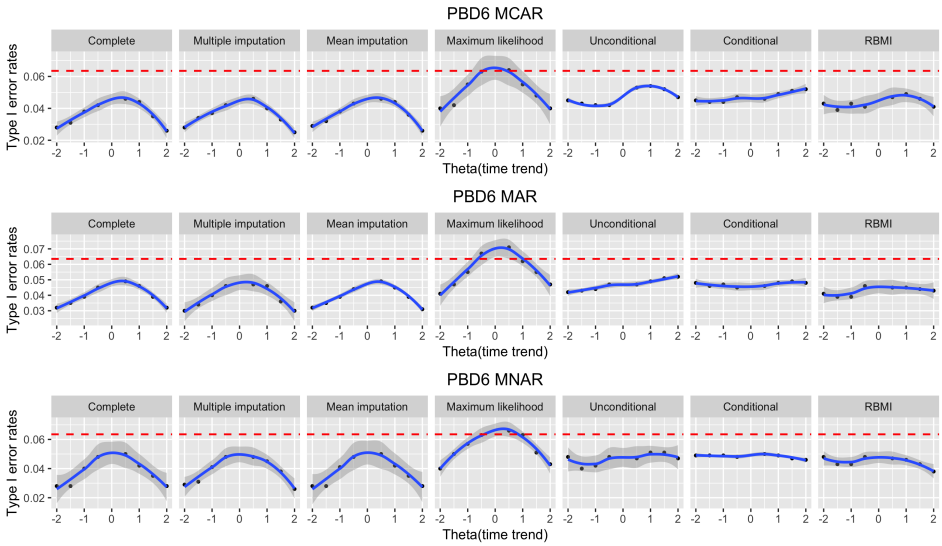

Figure A.63: *Time trend: PBD (blocksize = 6), type I error rates,* $p_{ms} = 0.1$*,* $n = 50$

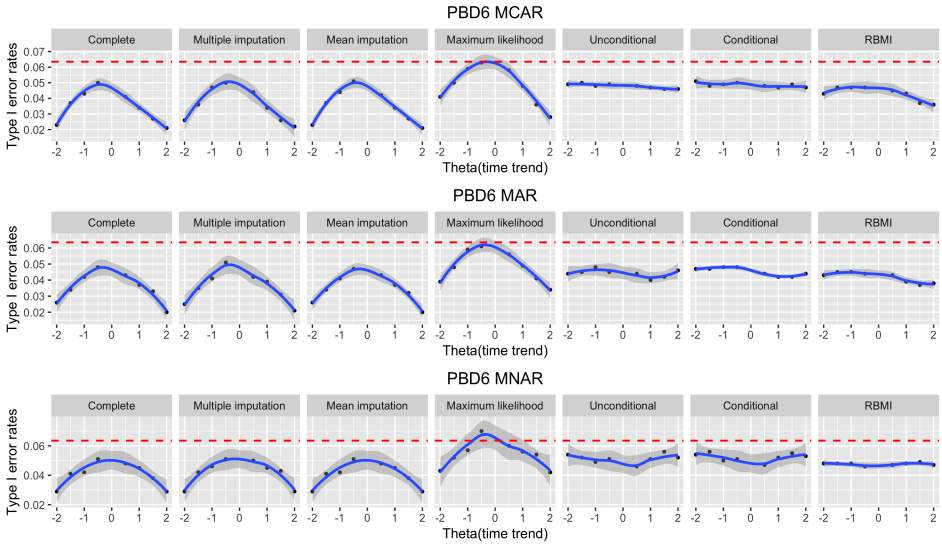

Figure A.64: *Time trend: PBD (blocksize = 6), type I error rates,* $p_{ms} = 0.1$, $n = 100$

A.4 Results Under Heterogeneity (Outliers)

A.4.1 Simulation Results under BCD

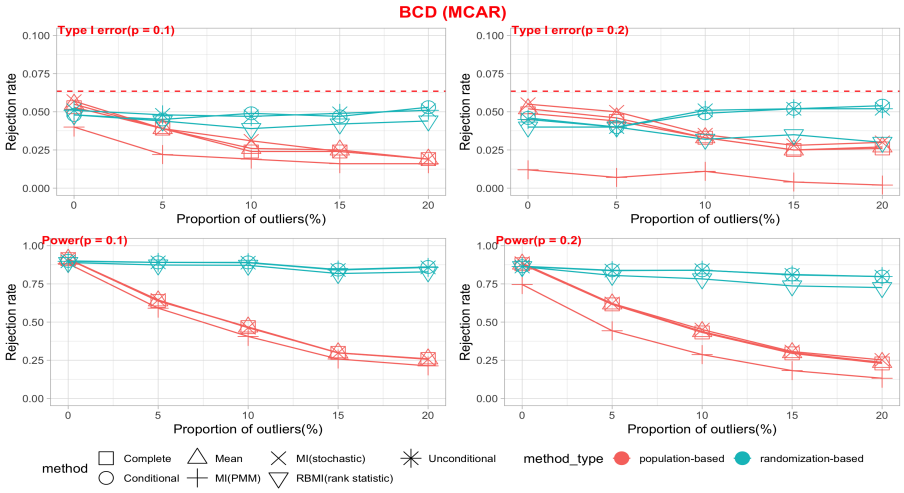

Figure A.65: *Outliers: BCD, MCAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

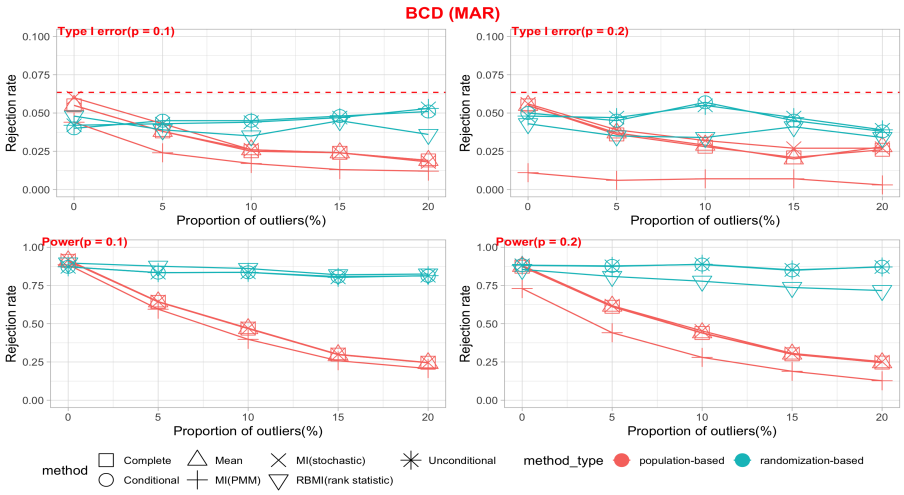

Figure A.66: *Outliers: BCD, MAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

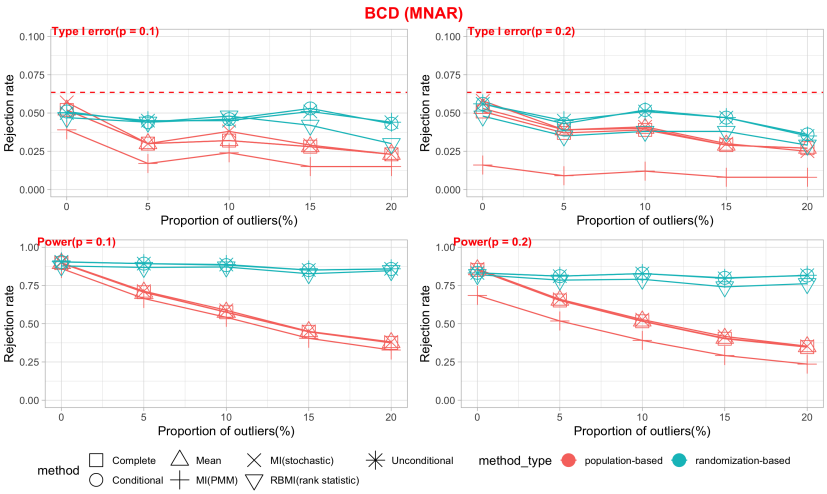

Figure A.67: *Outliers: BCD, MNAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

A.4.2 Simulation Results Under RBD

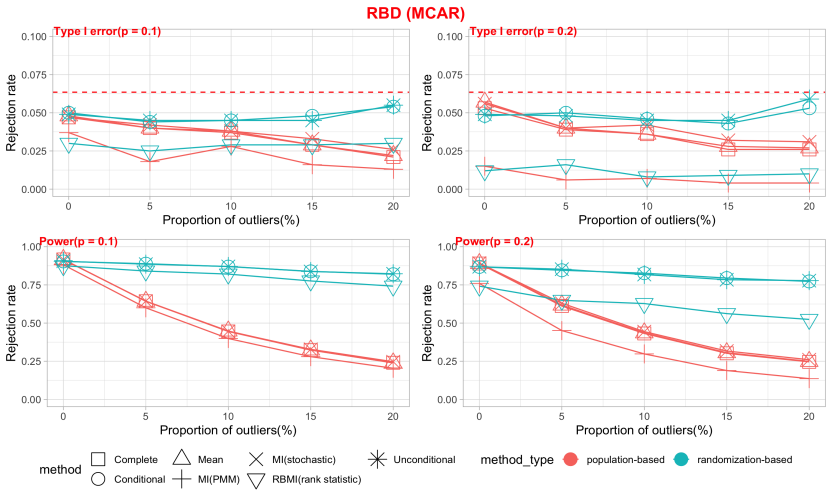

Figure A.68: *Outliers: RBD, MCAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

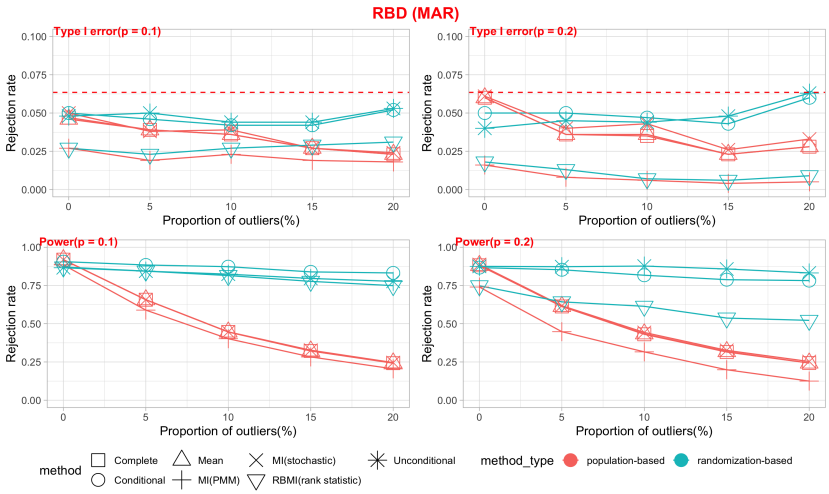

Figure A.69: *Outliers: RBD, MAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

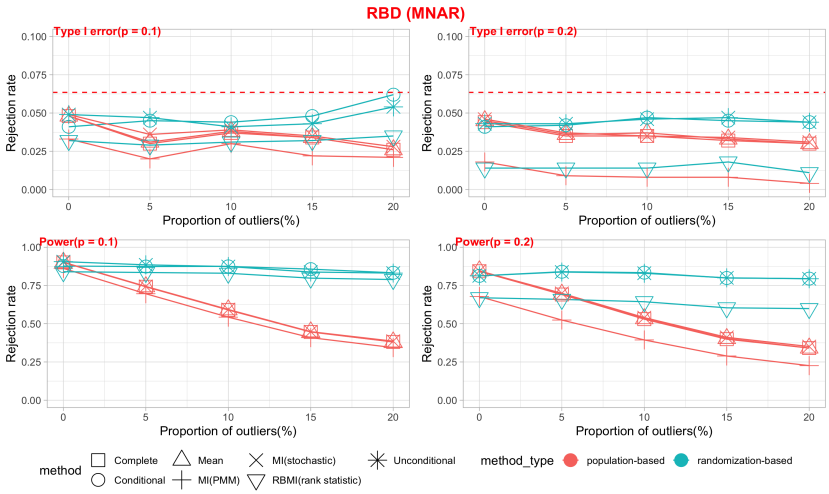

Figure A.70: *Outliers: RBD, MNAR, power & type I error rates,* $p_{ms} = 0.1$ *(left plots),* $p_{ms} = 0.2$ *(right plots),* $n = 100$

References

Blackwell, D. and Hodges, J. L. (1957). Design for the control of selection bias. *Annals of Mathematical Statistics*, 28:449–460.

Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.*

Calaway, R., Weston, S., Tenenbaum, D., and Calaway, M. R. (2015). Package 'doparallel'.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Bates, D., Chambers, J., and Eddelbuettel, M. D. (2021). Package 'rcpp'.

Edgington, E. and Onghena, P. (2007). *Randomization Tests.* CRC Press, New York.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58:403–417.

Enders, C. K. and Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8:430–457.

Galbete, A. and Rosenberger, W. F. (2016). On the use of randomization tests following adaptive designs. *Journal of Biopharmaceutical Statistics*, 26:466–474.

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics*, 52:1387–1393.

Kennes, L. N., Hilgers, R.-D., and Heussen, N. (2012). Choice of the reference set in a randomization test based on linear ranks in the presence of missing values. *Communications in Statistics-Simulation and Computation*, 41:1051–1061.

Lachin, J. M. (1999). Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials*, 20:408–422.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Wiley, New York.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer, New York.

Novo, A. A. and Schafer, J. L. (2013). *norm: Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.5.

Parhat, P., Rosenberger, W. F., and Diao, G. (2014). Conditional monte carlo randomization tests for regression models. *Statistics in Medicine*, 33:3078–3088.

Rosenberger, W. F. and Lachin, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*. Wiley, New York.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, 38:1–12.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press, New York.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7:147.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88:2909–2930.

Soares, J. F. and Wu, C. (1983). Some restricted randomization rules in sequential designs. *Communications in Statistics-Theory and Methods*, 12:2017–2034.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC press, New York.

Van Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–68.

Von Hippel, P. T. (2007). Regression with missing ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37:83–117.

Wang, Y. and Rosenberger, W. F. (2020). Randomization-based interval estimation in randomized clinical trials. *Statistics in Medicine*, 39:2843–2854.

Wang, Y., Rosenberger, W. F., and Uschner, D. (2020). Randomization tests for multiarmed randomized clinical trials. *Statistics in Medicine*, 39:494–509.

White, I. R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, 54:2267–2275.

Curriculum Vitae

Xiao Tan graduated from Harbin University of Commerce in 2015 with a B.S. degree in statistics. She received her M.A. degree in statistics from the University of California, Santa Barbara in 2017. She joined the Ph.D program in statistics at George Mason University in 2018, and worked as a Graduate Research Assistant at the Department of Statistics from 2018 to 2022.