2018-08-31

Treatment Effect Models for Subgroup Analysis with Missing Data

Fu, Yunting

UNIVERSITY OF CALGARY

Treatment Effect Models for Subgroup Analysis with Missing Data

by

Yunting Fu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

AUGUST, 2018

Abstract

The need for subgroup analysis in clinical trials in various contexts is increasing and data-driven approaches for subgroup identification based on statistical principles are desired. Among all subgroup identification methods, we focus on the treatment effect models that estimate the treatment contrast, since these models are intuitive and useful to interpretation. We evaluate and address the consequences of having missing data when using the Interaction Trees (IT), Qualitative Interaction Trees (QUINT) and Subgroup Identification based on Differential Effect Search (SIDES) methods. Simulation studies are used to demonstrate the accuracy of variable selection and bias in treatment effects when using complete, incomplete and imputed data across various scenarios when the sample size, proportion of missingness and imputation methods differ. We also applied these methods to a non-small cell lung cancer (NSCLC) dataset obtained from a retrospective study. Our results indicate that both IT and QUINT methods work equivalently well in most situations, while the SIDES results are, in general, less comparable due to the different mechanisms of the methods. The treatment effect models should be chosen based on the objective of the study, the sample size, the number of variables containing missing data, and the data structure. In terms of the methods for addressing missing data, an assumption of the data structure needs to be made during the method selection. MissForest is an excellent choice for a dataset with a tree-based structure, while MI methods would be a good fit for the other situations.

Preface

This thesis is an original work by the author. No part of this thesis has been previously published.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Dr. Karen Kopciuk and Dr. Hua Shen, for their continuous support of my study and related research. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better supervisors for my graduate study.

My sincere thanks also go to Dr. Ying Yan, my former supervisor, who provided me an opportunity to start this study. Without his previous support it would not be possible to conduct this research.

Besides my supervisors, I would like to thank the rest of my thesis committee: Dr. Alexander de Leon and Dr. Tolulope Sajobi, for their insightful comments.

I would also like to acknowledge Dr. Gwyn Bebb and the Glans-Look database. Without their kind support, the application study could not have been successfully conducted.

Last but not the least, I would like to thank my family: my parents and my brother, for supporting me spiritually throughout writing this thesis and my life in general.

Table of Contents

Abstract ii

Preface iii

Acknowledgements iv

Table of Contents v

List of Figures and Illustrations viii

List of Tables ix

List of Symbols, Abbreviations and Nomenclature x

1 Introduction 1
- 1.1 Heterogeneity in Treatment Effect . . . 2
- 1.2 Classification of Subgroup Analysis Studies . . . 4
- 1.3 Data-driven Approaches . . . 4
- 1.4 Scope of this Thesis . . . 6

2 Background 7
- 2.1 Subgroup Analysis Methods . . . 7
  - 2.1.1 Two Categories of Subgroup Analysis Models . . . 8
  - 2.1.2 General Setting . . . 9
- 2.2 Tree-based Partitioning Methods . . . 10
  - 2.2.1 Terminology . . . 11
  - 2.2.2 Splitting Algorithm . . . 11
  - 2.2.3 Pruning Algorithm . . . 13
- 2.3 Missing Data . . . 15
  - 2.3.1 Missing Mechanisms . . . 16
  - 2.3.2 Missing Data Patterns . . . 16

3 Methods 18
- 3.1 Interaction Trees . . . 18
  - 3.1.1 Growing a Large Tree . . . 18
  - 3.1.2 Pruning the Tree . . . 20

3.1.3 Determining the Best Subtree . . . 20
3.2 Qualitative Interaction Trees . . . 21
3.2.1 Splitting Criterion . . . 21
3.2.2 Pruning Algorithm . . . 23
3.3 Subgroup Identification based on Differential Effect Search . . . 24
3.3.1 Splitting Criterion . . . 25
3.3.2 Continuation Criterion . . . 26
3.3.3 Selection Criterion . . . 27
3.4 Handling Missing Data . . . 28
3.4.1 Single Imputation . . . 29
3.4.2 Multiple Imputation . . . 29

4 Simulation Study **31**
4.1 Study Design . . . 31
4.2 Model Performance . . . 33
4.3 Results . . . 34
4.3.1 Scenario One . . . 34
4.3.2 Scenario Two . . . 37
4.3.3 Scenario Three . . . 39
4.3.4 Scenario Four . . . 41
4.3.5 Scenario five . . . 43
4.3.6 Scenario Six . . . 45
4.3.7 Scenario Seven . . . 46
4.3.8 Scenario Eight . . . 48
4.4 Conclusions . . . 50

5 Application Study **60**
5.1 Study Design . . . 60
5.2 Parameters for the Methods . . . 62
5.3 Results . . . 63
5.3.1 The First Treatment Condition . . . 63
5.3.2 The Second Treatment Condition . . . 65
5.4 Conclusions . . . 67

6 Conclusions and Future Work **69**
6.1 Conclusions . . . 69
6.2 Future work . . . 70

Bibliography **72**

A Appendix for the Simulation Study **78**

B Appendix for the Application Study **82**
B.1 Results for the Mid-stage Outcomes . . . 82
B.1.1 Progression . . . 83
B.1.2 Recurrence . . . 84

B.2 Results for the End-stage Outcome 87
B.2.1 Vital Status 87

List of Figures and Illustrations

2.1 A Simple Tree Structure 11

4.1 Model 1: Tree Structure with Continuous Outcomes 33
4.2 The Frequency of Variable Selection for the First Scenario [1] 36
4.3 The Bias and Relative Bias for the First Scenario [2] 36
4.4 The Frequency of Variable Selection for Scenario Two [3] 38
4.5 The Bias and Relative Bias for Scenario Two [4] 38
4.6 The Frequency of Variable Selection for Scenario Three [5] 40
4.7 The Bias and Relative Bias for Scenario Three [6] 40
4.8 The Frequency of Variable Selection for Scenario Four [7] 42
4.9 The Bias and Relative Bias for Scenario Four [8] 42
4.10 The Frequency of Variable Selection for Scenario Five [9] 44
4.11 The Bias and Relative Bias for Scenario Five [10] 44
4.12 The Frequency of Variable Selection for Scenario Six [11] 45
4.13 The Bias and Relative Bias for Scenario Six [12] 46
4.14 The Frequency of Variable Selection for Scenario Seven [13] 47
4.15 The Bias and Relative Bias for Scenario Seven [14] 48
4.16 The Frequency of Variable Selection for Scenario Eight [15] 49
4.17 The Bias and Relative Bias for Scenario Eight [16] 49

A.1 The Bias and Relative Bias for Scenario One [17] 78
A.2 The Bias and Relative Bias for Scenario Two [18] 79
A.3 The Bias and Relative Bias for Scenario Five [19] 79
A.4 The Bias and Relative Bias for Scenario Six [20] 80
A.5 The Bias and Relative Bias for Scenario Six [21] 81

List of Tables

4.1 Detailed Information for Eight Scenarios 32
4.2 Performance Comparisons for Scenario One Based on 100 Simulations 52
4.3 Performance Comparisons for Scenario Two Based on 100 Simulations 53
4.4 Performance Comparisons for Scenario Three Based on 100 Simulations . . . 54
4.5 Performance Comparisons for Scenario Four Based on 100 Simulations 55
4.6 Performance Comparisons for Scenario Five Based on 100 Simulations 56
4.7 Performance Comparisons for Scenario Six Based on 100 Simulations 57
4.8 Performance Comparisons for Scenario Seven Based on 100 Simulations . . . 58
4.9 Performance Comparisons for Scenario Eight Based on 100 Simulations . . . 59

5.1 Candidate Covariates in NSCLC Study 61
5.2 Patient Characteristics in Two Treatment Groups: Survival Month Outcome with T_1 64
5.3 Results for the First Treatment Condition 64
5.3 Results for the First Treatment Condition 65
5.4 Patient Characteristics in Two Treatment Groups: Survival Month Outcome with T_2 66
5.5 Results for the Second Treatment Condition 66
5.5 Results for the Second Treatment Condition 67

B.1 Patient Characteristics in Two Treatment Groups: Mid-stage Data with T_2 . 83
B.1 Patient Characteristics in Two Treatment Groups: Mid-stage Data with T_2 . 84
B.2 Results for the SIDES: Progression Outcome under T_2 84
B.3 Patient Characteristics in Two Treatment Groups: Progression Outcome with T_1 85
B.3 Patient Characteristics in Two Treatment Groups: Progression Outcome with T_1 86
B.4 Results for the SIDES: Recurrence Outcome 86
B.5 Patient Characteristics in Two Treatment Groups: Vital Status Outcome with T_1 88
B.6 Results for the SIDES: Vital Status Outcome 88
B.6 Results for the SIDES: Vital Status Outcome 89

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
1-SE	1 Standard Error
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CART	Classification and Regression Trees
EGFR	Epidermal Growth Factor Receptor
EM Algorithm	Expectation-Maximization Algorithm
FIML	Full Information Maximum-Likelihood
GM	Genetic Matching
GUIDE	Generalized Unbiased Interaction Detection and Estimation
IPTW	Inverse Probability of Treatment Weighting
IT	Interaction Trees
MAR	Missing At Random
MCAR	Missing Completely At Random
MF	MissForest
MI	Multiple Imputation
MIPMM	Multiple Imputation based on Predictive Mean Matching
MIEM	Multiple Imputation based on Expectation-Maximization
MNAR	Missing Not At Random
NSCLC	Non-Small Cell Lung Cancer
OWL	Outcome-weighted Learning
PLRT	Partial Likelihood Ratio Test
PS	Propensity Score
QUINT	Qualitative Interaction Trees
RCT	Randomized Clinical Trials
RSS	Residual Sum of Squares
ROWSi	Regularized Outcome Weighted Subgroup Identification
SIDES	Subgroup Identification based on Differential Effect Search
U of C	University of Calgary

Chapter 1

Introduction

Subgroup analysis is a type of method developed to identify and confirm subgroups where a treatment is effective (Ondra et al., 2016). The increased need for subgroup analysis in randomized clinical trials (RCTs) in various contexts has been mentioned in many studies (Brookes et al., 2001; Rothwell, 2005; Mayer et al., 2015; Lipkovich et al., 2017). RCTs are one type of well-developed study in medical research, and they are considered to be the gold standard in assessing the safety and efficacy of medical interventions (Yusuf et al., 1984; Collins, 1996). Given the fact that most clinical trials are conducted at multi-centres and are time-consuming, they require a great amount of effort and resources. Because of the major investment involved in these studies, investigators normally attempt to extract more information from a successful clinical trial or sometimes try to save a failed trial. The best information we can gain from these studies is to identify patients who are more prone to benefit from the treatments than others. Patients in clinical trials vary in their background characteristics, such as demographic factors, genetic makeup, disease severity and status, which may affect their responses to certain treatments. Studies show that many patients do not respond to even commonly used treatments as their doctors expected (Ruberg and Shen, 2015). It is important to select the right treatment for the right groups of patients by evaluating the differences in treatment effect among subgroups, and this information provides

important evidence for personalized medicine (Goldberger and Buxton, 2013; Lipkovich et al., 2017). Thus, subgroup identification approaches based on statistical principles are desired (Pocock et al., 2002; Rothwell, 2005; Mayer et al., 2015; Lipkovich et al., 2017).

The main objective of subgroup analysis is to identify heterogeneity in treatment effect (Alosh et al., 2017; Lipkovich et al., 2017). Under the personalized medicine frameworks, the subgroup analysis procedures can be developed in two ways: (1) identify the right patient for a certain treatment; and (2) identify the right treatment for a particular patient, which is also referred as the tailored therapeutics (Ruberg and Shen, 2015; Lipkovich et al., 2017). The first approach aims to identify the magnitude in the differential treatment effect, while the second approach focuses on the direction of the treatment difference (Lipkovich et al., 2017). In this process, different studies need different types of subgroup analysis approaches for their methodology. Detailed information is provided in the following sections on these methods.

1.1 Heterogeneity in Treatment Effect

Subgroup analysis is a method evaluating the treatment effect in subgroups of patients and focuses on identifying predictive markers (Wang et al., 2007). A predictive marker is a clinical or biologic characteristic that provides information on the outcome of interest, under the effect of a particular treatment (Rothwell, 2005; Italiano, 2011). In contrast to predictive markers, a prognostic marker is a factor that provides information on the event of interest in the absence of any treatment. Prognostic markers can be used to adjust a relative treatment effect to a absolute scale and provide further information on new interventions (Hemingway et al., 2013; Riley et al., 2013). While both are important in clinical trials, predictive markers are commonly of interest in subgroup analysis to estimate the treatment differences among subgroups.

One challenge in identifying predictive markers is to successfully investigate the hetero-

geneity of treatment effects across subgroups. Heterogeneity can be classified as quantitative and qualitative treatment effects (Wang et al., 2007). Quantitative treatment effects are those where the treatment effects vary in magnitude by subgroups, but always in the same direction. In other words, one treatment is always better than another one, while qualitative treatment effects represent those where the directions of the treatment effect are different across subgroups. For example, in the first subgroup, the experimental treatment is better than the control treatment, while in the second subgroup, the control treatment is better than the experimental one.

The best known statistical method for detecting the heterogeneity of treatment effects across subgroups is testing for treatment-covariate interactions (Wang et al., 2007). A common mistake in evaluating the treatment effect is considering the heterogeneity in patients as the heterogeneity in treatment effect (Wang et al., 2007). Either a crossover clinical trails design or a evaluation of treatment-covariate interactions can be used to address the problem. Consider a dataset which is obtained from a non-crossover study, testing for treatment-covariate interactions will be the first option to detect the heterogeneity of treatment effects. Furthermore, testing for a quantitative interaction in a two-arm study is relatively easy, since a standard method not only leads to a conclusion about the existence of an interaction, but also provides information related to the direction. However, a statistically significant result for a qualitative interaction is not sufficient enough to determine which group receives a greater effect from the experimental treatment and which group benefits more from the control treatment (Alosh et al., 2015). Further testing is needed and the treatment effects within subgroups should be reported. Successfully identifying qualitative interactions might be vital for salvaging a new potentially effective treatment from a clinical trial with a non-significant primary outcome (Alemayehu et al., 2017).

1.2 Classification of Subgroup Analysis Studies

By evaluating a differential treatment effect, we are able to give strong evidence to confirm the presence of subgroups or to develop personalized medicine. Based on the different contexts and objects of subgroup analysis, Mayer et al. (2015) classified them into four types of studies: confirmatory subgroup analysis, exploratory subgroup analysis, post hoc and subgroup discovery.

We can consider confirmatory subgroup analysis as one type of hypothesis testing. It aims at evaluating a small number of pre-defined subgroups which are identified in the earlier stage of studies. In contrast, exploratory subgroup analysis is more related to the idea of hypothesis generation or enrichment strategies (Mayer et al., 2015; Lipkovich et al., 2017). This analysis strategy is normally specified at the protocol development stage and is used to assess subgroups by considering the treatment-covariate interactions. Post hoc subgroup evaluation is performed to address situations or issues that arise after the main studies results are available, such as safety issues, post-marketing activities, and heterogeneity in multi-regional studies. Subgroup discovery study employs data mining and machine learning techniques to investigate a large pool of candidate subgroups and provide useful information for future trials (Mayer et al., 2015; Lipkovich et al., 2017).

1.3 Data-driven Approaches

All three types of subgroup analysis, except for the confirmatory one, require data-driven elements for their statistical methodology (Lipkovich et al., 2017). In contrast to data-driven approaches, traditional guideline-driven methods require criteria such as *a priori* specification of the subgroups and the subgroup effects, consistency of the subgroup effect across outcomes, and pre-specification of the directions of subgroup effects (Sun et al., 2010). Some of these criteria run counter to the very spirit of the subgroup identification and may lead to the insufficient use of data. Lipkovich et al. (2017) believe that data-driven approaches, as a

special case of model selection, can obtain more information from the data compared to the guideline-driven methods.

Data-driven approaches address the uncertainty related to the subgroup specification. As a methodology developed from machine/statistical learning techniques, data-driven approaches can identify a small number of subgroups from a large pool of candidate subgroups without any pre-specification requirement (Lipkovich et al., 2017). Similar to general machine learning techniques, data-driven approaches can be either generated from the regression-based methods or from other arbitrary methods such as the tree-based methods. On account of their intuitive algorithm, tree-based methods normally lead to an objective and easily interpretable result.

Setting aside all the advantages, data-driven approaches also share some common restrictions with guideline-driven approaches, such as multiplicity, data over-fitting, and selection bias (Lipkovich et al., 2017). Fortunately, different types of methods are available for addressing those problems. We can apply re-sampling methods (e.g. bootstrap and cross-validation) to adjust for multiplicity. Complexity controlling methods, such as pruning, can be applied to prevent over-fitting the data. Selection bias is a major concern in data-driven approaches. Because different covariates have different numbers of possible splits, covariates with larger numbers of possible splits will have more chance to be chosen compared to the ones with smaller numbers of possible splits (Lipkovich et al., 2017). There are also methods for handling selection bias. For example, Generalized Unbiased Interaction Detection and Estimation (GUIDE) selects the best covariates using a simple test statistic to adjust for the possible cutoffs of the predictor values (Loh, 2002).

Given the irresistible benefits and solvable limitations, data-driven approaches play an important part in subgroup analysis, especially for studies such as exploratory subgroup analysis, post hoc and subgroup discovery. Our study focuses on three tree-based data-driven methods, which are the Interaction Trees (IT), Qualitative Interaction Trees (QUINT) and Subgroup Identification based on Differential Effect Search (SIDES), which will be introduced

in details in Chapter 3.

1.4 Scope of this Thesis

In this thesis, we evaluate and address the consequences of having missing data when using the IT, QUINT and SIDES methods. All three methods are tree-based partitioning approaches that evaluate treatment-covariate interactions.

The remainder of this thesis is structured as follows: Chapter 2 provides detailed information on subgroup analysis models, tree-based partitioning methods and missing data patterns/mechanisms. Chapter 3 introduces the IT, QUINT, SIDES and the methods we used to address the missing data. The design and results of the simulation study are introduced in Chapter 4. Chapter 5 introduces the real data application of our methods. The final chapter summarizes our work and also formulates some future works.

Chapter 2

Background

In this chapter, we introduce the general setting of subgroup analysis methods, algorithms of tree-based methods, and properties of missing data. In the previous chapter, we mentioned that most subgroup analysis studies require data-driven elements in their statistical methodologies, and tree-based methods are one of the most intuitive and objective data-driven methods. Most of these tree-based approaches introduce some ad-hoc methods to address missing data. Given the importance of handling missing data in real-life applications, we evaluated the performance of these tree-based approaches when applying their default methods and other data completion methods to address missing values.

2.1 Subgroup Analysis Methods

Based on the Mayer et al. (2015) definition of subgroup analysis studies, data-driven approaches are essential for three types of studies: exploratory subgroup analysis, post hoc and biomarker discovery. Lipkovich et al. (2017) categorized the data-driven subgroup evaluation methods as global outcome modeling, global treatment effect modeling, optimal treatment regimes and local modeling methods. Among these four classifications, global outcome modeling methods evaluate the underlying outcome function, while the global treatment effect models estimate the treatment contrast. Optimal treatment regimes are a special case of

global treatment effect modeling, which aim at identifying the right treatment for a set of patient's covariates. Local modeling methods deal with the direct search of subgroups who have a better treatment benefit.

2.1.1 Two Categories of Subgroup Analysis Models

According to the classifications defined by Lipkovich et al. (2017), we recategorize the data-driven subgroup evaluation methods into two groups: treatment effect modeling methods and outcome modeling methods. While the four classifications might have advantages by comprehensively considering the methodology and detailed differences in concepts among all methods, we are mainly focusing on the statistical principles. Thus, we opt to use the two large categories for convenience. We combine the global treatment effect modeling, optimal treatment regimes and local modeling methods into one group: the treatment effect methods, since their boundaries are blurry and there is some overlap in their underlying statistical methodologies.

For the treatment effect methods, what we aim to obtain is subgroups with maximum heterogeneity in their treatment effects. These methods focus on evaluating the treatment contrast $z(\boldsymbol{X})$ instead of the outcome function $f(\boldsymbol{X}, T)$, where $\boldsymbol{X}$ and T represent the vectors of covariates and treatments respectively. Optimal treatment regimes, which emphasize identifying the best treatment for a given patient, are a special case of global treatment effect modeling (Lipkovich and Dmitrienko, 2014a). It is reasonable to categorize the optimal treatment regimes into our treatment effect methods for simplification. Local modeling methods are approaches searching for subgroups with a better treatment benefit, in other words, subgroups with a higher value of $z(\boldsymbol{X})$ (Lipkovich et al., 2017). The main difference between the global treatment effect models and the local modeling methods is their searching scale. The global treatment effect models tend to discover several subgroups while the local modeling methods aim at identifying one unique subgroup. Without regard to the searching scale, both models belong to the treatment effect methods.

The outcome modeling methods evaluate the outcome function $f(\boldsymbol{X}, T)$. This implies that if there are a set of observed covariates, the expected outcome $\hat{Y}$ can be estimated by fitting the covariates to the model obtained from these approaches. These types of methods are well-developed in statistics and have been applied in different fields. The following section shows the general setting of the treatment effect and the outcome modeling methods.

2.1.2 General Setting

In a two-arm clinical trial setting, we define Y as the outcome variable and y_i as the response for the i^{th} patient ($i = 1, 2, \ldots, n$). Similarly, $\boldsymbol{X}$ is the vectors of covariates and let $\{x_{i1}, \ldots, x_{ip}\}$ denote a vector of observed biomarkers for the i^{th} patient. Let T denote the treatment assignment and let t_i be the treatment for i^{th} patient, where t_i=0 for the control treatment group and t_i=1 for the experimental treatment group. Lipkovich et al. (2017) defined the expected response of a patient and the treatment contrast at the patient's level as $f(\boldsymbol{X}, T) = E(Y|\boldsymbol{X}, T)$ and $z(\boldsymbol{X})$, respectively. The relationship between the $f(\boldsymbol{X}, T)$ and $z(\boldsymbol{X})$ is shown as follows. The general formula of $z(\boldsymbol{X})$ is defined in Lipkovich et al. (2017) as:

$$z(\boldsymbol{X}) = h\left(f(\boldsymbol{X}, 1), f(\boldsymbol{X}, 0)\right),$$

where $h(x)$ is a monotone function. If the outcome is a continuous variable, then $z(\boldsymbol{X})$ is equal to:

$$z(\boldsymbol{X}) = f(\boldsymbol{X}, 1) - f(\boldsymbol{X}, 0),$$

which leads to $f(\boldsymbol{X}, T) = f(\boldsymbol{X}, 0) + z(\boldsymbol{X})T$. In general, the outcome function is modelled as:

$$f(\boldsymbol{X}, T) = h\left(\ell(\boldsymbol{X}) + g(z(\boldsymbol{X})T)\right),$$

where $\ell(\boldsymbol{X})$ is a baseline function and $g(x)$ is a monotone function (Xu et al., 2015).

Based on this representation, it is easy to understand the difference and the relationship between the outcome modeling methods and treatment effect modeling methods we

employed in this thesis. Moreover, the last formula also shows the relationship between the prognostic markers and the predictive markers that we described in the first chapter. The prognostic markers only contribute to the baseline effect $\ell(\boldsymbol{X})$, while the predictive markers can also contribute to the treatment contrast $z(\boldsymbol{X})$ in the treatment-covariate interaction term (Lipkovich et al., 2017).

From these two categories, we will use the treatment effect models to assess the predictive effect as it provides information about a therapeutic intervention. The choice of treatment effect methods can be either based on a regression framework or a tree structure. Regression-based approaches are normally based on fitting a penalized regression model, such as the outcome-weighted learning (OWL) method and the regularized outcome weighted subgroup identification (ROWSi) (Zhao et al., 2012; Xu et al., 2015). Tree-based approaches, however, are more arbitrary, dividing the population into subgroups without any parameter estimation. All three methods we introduce in this thesis are tree-based treatment effect models (Su et al., 2009; Lipkovich et al., 2011; Dusseldorp and Van Mechelen, 2014). The following section introduces the general algorithms of the tree-based methods.

2.2 Tree-based Partitioning Methods

Tree-based partitioning methods are nonparametric machine learning approaches which have been widely used in many scientific and commercial areas (Strobl et al., 2009; Loh, 2011). The most well-known tree-based partitioning method is the classification and regression trees (CART) proposed by Breiman et al. (1984). Many new methods have been developed based on the idea of CART (Zhang and Singer, 2010). In this section, we introduce the general setting of most tree-based partitioning methods, instead of focusing on any particular one.

2.2.1 Terminology

Figure 2.1 shows a simple tree structure with three layers: a root node, internal nodes, and terminal nodes. A root node is the top node, and there must be only one root node in any tree. An internal node is the one with at least one branch. Hence, the root node can also be considered as an internal node (Zhang and Singer, 2010). The leaves of a tree are called as terminal nodes. In a tree structure, a node that has branches from it is also called as a parent node, while the branch is called as a child node. The aim of a tree-based method is to obtain relatively homogeneous terminal nodes (Zhang and Singer, 2010). The algorithms adopted to fulfill the goal are covered in following sections.

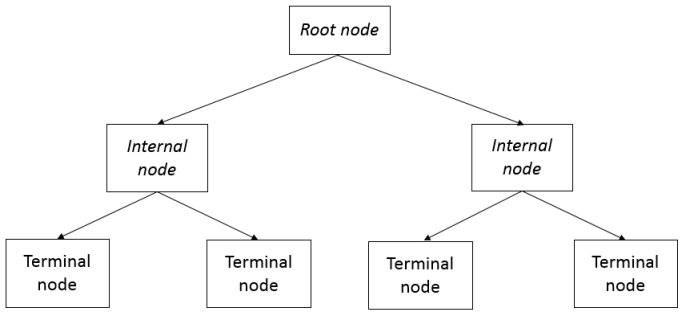

Figure 2.1: A Simple Tree Structure

2.2.2 Splitting Algorithm

Classification and regression trees share the same underlying concept, which is dividing the predictor space into J distinct and non-overlapping regions (or nodes) (1, 2, ..., J). For a binary split, the best split is the one that results in two nodes which are homogeneous in themselves and heterogeneous between each other. Therefore, we need to measure the node impurity to ensure the homogeneity of each node (Zhang and Singer, 2010). There are different measurements for node impurity, depending on the properties of the outcome variables for the classification and regression trees.

For regression trees, the residual sum of squares (RSS) is adopted to estimate the impurity. The RSS is defined as (James et al., 2013):

$$RSS = \sum_{j=1}^{J} \sum_{i \in j} (y_i - \hat{y}_j)^2,$$

where $\hat{y}_j$ is the average of y_i in node j. At each split, we choose the best combination of the splitting variable and the cutpoint with the smallest within-node variance. Once a split is finalized, the process moves down to the second level and successive splits by minimizing the RSS. The algorithm is repeated until a stopping criterion, such as reaching a pre-set minimum number of observations for each terminal node, is met.

For classification trees, we consider a subgroup to be pure if there is only one single class. The possibility for a subgroup to contain more than one class is measured by the impurity function. Common choices of the impurity function for a classification tree is one of the following: the misclassification error rate, Gini index and entropy (James et al., 2013). Since the category with the largest proportion of observations is denoted as the true classification for a subgroup, the misclassification error rate for k classes is defined as:

$$E = 1 - \max_k(\hat{p}_{jk}),$$

where $\hat{p}_{jk}$ represents the proportion of observations in the j^{th} node that are from the k^{th} class. The Gini index is a measure of total variance across the classes. It takes into account of both the probabilities that an observation is categorized correctly ($\hat{p}_{jk}$) and incorrectly ($1 - \hat{p}_{jk}$). The definition of Gini index is shown as follows:

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk}).$$

Entropy is measuring the mean information content of a certain event x. The information content for x is defined as $I(x) = -\log(p(x))$, where $p(x)$ denotes the probability of occurring.

The higher the value of p, the less information content we obtain. As the expectation of the information content, entropy is defined as:

$$D = -\sum_{k=1}^{K} \hat{p}_{jk} \log \hat{p}_{jk}.$$

Entropy for any pure subgroup is equal to zero, since $\hat{p}_{jk}$ is 1 and $\log \hat{p}_{jk}$ is 0. For any subgroup where all classes have equal probabilities, entropy reaches its highest value.

All impurity functions share similar properties: they achieve a minimum value when $\hat{p}_{jk}$ is 1 while achieving a maximum value when the probabilities of all classes are equal, and are symmetric with regard to their permutations. However, the misclassification error rate is rarely used due to its insensitivity to the variance involved with tree-growing. The other two impurity functions are preferred in practice and their numerical results are generally quite similar (Zhang and Singer, 2010; James et al., 2013). Similar to the regression trees, the classification trees choose the split with the smallest value of the impurity function at each step, and repeat this algorithm successively from the root node to the terminal node.

After applying such a top-down splitting algorithm, over-fitting is a challenging problem that needs to be addressed. Instead of introducing a stopping rule, the widely accepted solution for over-fitting in tree-based methods is to grow a large tree first, and then prune it back to a smaller size. The following section covers the algorithm in the pruning step.

2.2.3 Pruning Algorithm

The problem for a large tree is this: it might perform well with the original data, but is normally accompanied by large testing error when applied to a new dataset. The purpose of pruning is to achieve a balance between the number of splits and the testing error by finalizing a smaller subtree. In order to end up with the best subtree, we apply a cost-complexity function in the pruning procedure. First, we need to define the tree cost, $R(T)$, which is evaluating the overall misclassification error for all terminal nodes. Similar to the

splitting algorithm, the tree cost is equal to the residual sum of squares of all terminal nodes for the regression tree, and is commonly derived from the misclassification error for the classification trees (Zhang and Singer, 2010; James et al., 2013). Two tree costs are defined as follows:

For the regression tree:

$$R(T) = \sum_{m=1}^{|\tilde{T}|} \sum_{i \in m} (y_i - \hat{y}_m)^2,$$

where $|\tilde{T}|$ is the number of terminal nodes of the tree T and $\hat{y}_m$ is the average of y_i in the terminal node m. For the classification tree:

$$R(T) = \sum_{m=1}^{|\tilde{T}|} \left[\frac{n(m)}{n} (1 - \max_k(\hat{p}_{mk})) \right],$$

where $n(m)$ is the number of observations in terminal node m and n is the total number of the observations. $\hat{p}_{mk}$ represents the proportion of observations in the m^{th} terminal node that are from the k^{th} class.

One issue arising from using the tree cost to split the tree is that child nodes always have a smaller tree cost compared to the parent node, so it tends to bias toward a bigger tree. In order to prevent the tree from being over-fit, a penalty function is applied to make a trade-off between the tree size and the tree cost. The tree cost-complexity is defined as:

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}|,$$

where α ($\alpha \geq 0$) is the tuning parameter controlling the tree's complexity (Zhang and Singer, 2010). For any internal node, T, we cut all its offspring nodes when the cost-complexity of the internal node is smaller than that of the offspring terminal nodes, T_t. This procedure is also known as the weakest link pruning, and is defined as:

$$R(T) + \alpha \leq R(T_t) + \alpha |\tilde{T}_t|.$$

By calculating the $R(T)$ and $R(T_t)$, we can obtain the value of α for each internal node, which is:

$$\alpha = \frac{R(T) - R(T_t)}{|\tilde{T}_t| - 1}.$$

The smallest α is the first threshold parameter. If we have k thresholds and let $\alpha_0 = 0$, then for each $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k$, the corresponding subtrees will be $T_{\alpha_0} \supset T_{\alpha_1} \supset T_{\alpha_2} \supset \cdots \supset T_{\alpha_k}$ (Zhang and Singer, 2010). Normally, we use K-fold cross-validation to estimate the prediction error and then choose the best α with the smallest average error. The basic idea of K-fold cross-validation is that we randomly divide the dataset into K similar sized groups, and use the first group as a validation set and the remaining $K - 1$ groups as a model-fitting set. Breiman et al. (1984) also proposed a revised version of cross-validation for selecting the final subtree which is called the 1 standard error (1-SE) rule. Based on the 1-SE rule, we select the smallest subtree whose cross-validation estimate is within the one unit SE range of the α with the smallest average error.

Based on the algorithms of the tree-based partitioning methods, we introduce three new tree-based treatment effect methods in the next chapter, and also compare the model performance of these models when missing data is presented. The next section sheds some light on the properties of missing data.

2.3 Missing Data

In real-life studies, information is frequently lost due to several reasons. For example, participants in a cohort study may drop out of the study and there is no way to follow up on them. Lab results might be missing due to equipment errors. Regardless of the reason, missing data poses difficulties for data analysis and the lack of information may lead to incorrect conclusions. An incorrect conclusion drawn from a clinical study may result in an inappropriate treatment strategy and endanger a patient's life. Therefore, it is important to develop methods to deal with missing data. Before setting out to address the problem, it is

worth spending time to understand the missing mechanisms and missing data patterns.

2.3.1 Missing Mechanisms

There are three mechanisms behind missing data, which are missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR). For a data matrix $\boldsymbol{X}$ with two parts: the observed part ($\boldsymbol{X}_{obs}$) and the missing part ($\boldsymbol{X}_{mis}$), Little and Rubin (2014) defined MCAR as a condition in which the missingness does not depend on either the observed values, $\boldsymbol{X}_{obs}$, or the missing values, $\boldsymbol{X}_{mis}$. MAR is an assumption less restrictive than MCAR. It assumes that the missingness depends on the observed values, rather than the missing ones (Little and Rubin, 2014). The MNAR occurs when the missingness depends on the missing value itself, which makes it impossible to estimate.

MAR is the most common case in the real-life data, and there are many methods to address missing data, such as Multiple Imputation (MI), Full Information Maximum-Likelihood (FIML) and Expectation-Maximization (EM) algorithm. These methods are applicable under the MAR (together with MCAR) condition (Dong and Peng, 2013). Detailed information about MI and other imputation methods are covered in the next chapter.

2.3.2 Missing Data Patterns

Missing data patterns consider both the values that are observed and those that are missing. There are three types of missing patterns: univariate, monotone and arbitrary. If the same subjects have missing values on one or more of the variables, the missingness of that dataset is considered as a univariate missing pattern. The monotone missing pattern normally happens in longitudinal studies. For example, for a subject i at time j, when X_{ij} is missing, the following $X_{i,j+1}$, $X_{i,j+2}$, $\ldots$, X_{iK} will be missing too ($j = 1, \ldots, K$) (Dong and Peng, 2013). The arbitrary missing data pattern assumes that the missing data occurs at any unit randomly. It is relatively easy to handle the univariate and the monotone missing pattern, compared to the arbitrary missing pattern. In our study, we assume that the missing data

follows the univariate missing pattern.

Chapter 3

Methods

In this thesis, we evaluate and address the consequences of having missing data when using three treatment effect models: Interaction Trees (IT), Qualitative Interaction Trees (QUINT) and Subgroup Identification based on Differential Effect Search (SIDES). In this chapter, we introduce detailed information on these three models, and the methods that are applied to handle missing data.

3.1 Interaction Trees

Su et al. (2009) proposed a data-driven and automated partitioning procedure to identify the heterogeneity of the treatment effect among subgroups. The IT method is derived from the CART algorithm and shares similar splitting and pruning algorithms. This method consists of the following three fundamental steps: (1) growing a large tree; (2) pruning the tree back to a relatively smaller size; and (3) determining the best subtree.

3.1.1 Growing a Large Tree

For the IT method, the splitting criterion not only aims to generate pure nodes, but also maximizes the heterogeneity in treatment effect between two child nodes. Su et al. (2009)

proposed a simple t test to evaluate the heterogeneity. For each binary split, the test statistic is defined as:

$$t(s) = \frac{(\bar{y}_{left,T=1} - \bar{y}_{left,T=0}) - (\bar{y}_{right,T=1} - \bar{y}_{right,T=0})}{\hat{\sigma} \cdot \sqrt{1/n_1 + 1/n_2 + 1/n_3 + 1/n_4}},$$

where $\bar{y}$ and n are the sample means and the sample sizes, corresponding to each treatment group and child node. $\hat{\sigma}$ is the pooled estimate of the total sample variance.

The test for heterogeneity is equivalent to the t-test for testing $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$ in the following regression model:

$$y = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot Z + \beta_3 \cdot T \cdot Z + \epsilon,$$

where Z is the binary indicator variable used to split covariate X. If X is a continuous variable, $Z = \mathbb{1}\{X \leq c\}$ with a splitting point c, while if X is a categorical variable, $Z = \mathbb{1}\{X \in c\}$ where c is a subset of X. T is the treatment variable and $T \cdot Z$ is the interaction term between the treatment and the indicator variable, Z.

After testing all the possible splits, the best split s^* is the one with the largest interaction effect. In order to control the direction of the t statistics in the comparisons, the IT method uses the G statistic to compare the interaction effect instead of using the t statistics directly. The G statistic is defined as:

$$G(s) = t^2(s),$$

which follows a χ^2_1 distribution. We choose the best split s^* based on:

$$G(s^*) = \max G(s).$$

The splitting process stops when one of the following stopping rules is met: the pre-set minimum node size, the pre-specified maximum level of a tree, or the pre-set minimum number of observations for the child nodes (Su et al., 2009). For survival data, Negassa et al. (2005) and Su et al. (2008) proposed using the partial likelihood ratio test (PLRT) to assess

the interaction effect between subgroups. Detailed information can be found in these papers (Negassa et al., 2005; Su et al., 2008). No matter what kind of splitting algorithm is applied, once this process stops, we need a pruning algorithm to control the tree complexity.

3.1.2 Pruning the Tree

The pruning step aims to truncate the weakest node and then continuous to the next weakest node. At this step, the IT method adopts the split-complexity pruning algorithm proposed by LeBlanc and Crowley (1993). The difference between the general decision tree and the IT method is that instead of using the tree cost in the complexity function, the IT method introduces an interaction-complexity measure $G_\alpha(T)$. That is, for an initial large tree, T, with terminal nodes, $\tilde{T}$, the total number of internal nodes is defined as $|T-\tilde{T}|$, and $G(T) = \sum_{h\in(T-\tilde{T})} G(h)$ is the sum of the G statistic (measurement of the interaction effect) for all internal nodes. Su et al. (2009) defined the interaction-complexity as:

$$G_\alpha(T) = G(T) - \alpha \cdot |T - \tilde{T}|,$$

where α is the complexity parameter. By tuning α, we can end up with different tree sizes and different amounts of interaction effects. The main idea is to achieve the balance between the interaction effect $G(T)$ and the tree complexity $|T - \tilde{T}|$. The choice of the right α helps us obtain the best subtree and is described in following section.

3.1.3 Determining the Best Subtree

The main idea of determining the best subtree is to maximize G_α, and in the meanwhile to reduce the tree complexity. From Chapter 2.2.3, we know that for each $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k$, the corresponding subtrees are $T_{\alpha_0} \supset T_{\alpha_1} \supset T_{\alpha_2} \supset \cdots \supset T_{\alpha_k}$ (let $\alpha_0 = 0$). In order to finalize the best subtree, Su et al. (2009) suggested two ways to select α: (1) select arbitrary values from $2 \leq \alpha \leq 4$, where 2 is obtained from Akaike information criterion (AIC) and

4 is the approximate critical value for χ^2_1 when p is 0.05; or (2) use $\alpha = ln(n)$, which is derived from the Bayesian information criterion (BIC), where n is the total samples. These two approaches seem to provide similar results in terms of the tree size in our simulation study.

After finalizing the subtree, a validation method can be performed to assess the final tree. If the sample size is large enough, we can draw a subset from the data to assess the performance of the tree by calculating the predictive error. Otherwise, re-sampling techniques such as K-fold cross-validation or bootstrapping can be used as a validation method (LeBlanc and Crowley, 1993; Su et al., 2009). In practice, we normally use $K = 5$ or $K = 10$, which take into account a bias-variance trade-off (James et al., 2013).

3.2 Qualitative Interaction Trees

Dusseldorp and Van Mechelen (2014) introduced a sequential partitioning method to identify qualitative treatment-covariate interactions. Qualitative interaction means that the directions of treatment effect are different depending on the subgroups. By detecting qualitative interactions, the goal of the QUINT is to classify the population into three exclusive types of subgroups: (1) patients for whom the experimental treatment is better than the control treatment; (2) those for whom the control treatment is better than the experimental treatment; and (3) a group label that does not have any treatment difference.

3.2.1 Splitting Criterion

The splitting criterion of QUINT contains two parts: the difference in treatment outcome component and the cardinality component (Dusseldorp and Van Mechelen, 2014). The difference in treatment outcome component is the weighted average of the difference in treatment

outcome for all nodes in each subgroup, which can be quantified as:

$$D_c = \frac{\sum_{\ell=1}^{L} \left[\mathbb{1}(A(\ell) = c)\, N_{R_\ell}\, \alpha_\ell\, (-1)^c\, (\bar{y}_{T=1,\ell} - \bar{y}_{T=0,\ell}) \right]}{\sum_{\ell=1}^{L} \left[\mathbb{1}(A(\ell) = c)\, N_{R_\ell} \right]} \quad c = 1,\ 2,$$

where c represents the two subgroups with the treatment difference. $A(\ell)$ is an assignment function that assigns each leaf ℓ ($\ell = 1, \ldots, L$) to one of the partition subgroups. L is the total number of leaves. $\mathbb{1}(A(\ell) = c)$ is an indicator function testing if the leaf is in subgroup c. N_{R_ℓ} denotes the sample size in node R_ℓ and $\bar{y}_T$ is the sample mean of each subgroup corresponding to the different treatment assignments.

There are two different ways of calculating α_ℓ. If we aim at measuring the difference in means of outcome Y, then α_ℓ is simply equal to 1. If the effect size Cohen's d, the standardized difference between the two means corresponding to the different treatments, is evaluated, then $\alpha_\ell = 1/s_\ell$, with

$$s_\ell = \sqrt{\frac{(n_{T=0} - 1)\, s^2_{T=0,\ell} + (n_{T=1} - 1)\, s^2_{T=1,\ell}}{n_{T=0} + n_{T=1} - 2}},$$

where n and s^2 are the sample size and the sample variance corresponding to the different treatments in each subgroup. The choice of these two measurements relies on the scale of the outcome variable. If the outcome is a scale of values without any well-specified meaning, the effect size measurement is preferred. Otherwise, the difference in means should be used as the splitting criterion measurement (Dusseldorp et al., 2016).

The cardinality component is the sum of the number of patients in nodes assigned to each subgroup, which is defined as:

$$N_c = \sum_{\ell=1}^{L} \left[\mathbb{1}(A(\ell) = c)\, N_{R_\ell} \right] \quad c = 1,\ 2\,.$$

In order to maximum both components in practice, the QUINT method defines an overall partitioning criterion (C) by combining the difference in the treatment outcome component

and the cardinality component. That is:

$$C = \omega_1 \left[\log(1 + D_1) + \log(1 + D_2)\right] + \omega_2 \left[\log(N_1) + \log(N_2)\right],$$

where ω_1 and ω_2 denote the weights of the two components respectively. D_1 and D_2 are the differences in treatment outcome in the first and second subgroup, respectively. N_1 and N_2 are the cardinality components for the two subgroups. The log scale is used for convenience. Corresponding, we can easily add up all the components, instead of multiplying them.

We apply the partitioning algorithm to optimize the combination of the splitting point, the splitting variable and the number of leaves. By calculating the C values for all candidate nodes, the splitting point and the splitting variable with the highest C value are chosen. The algorithm stops when one of the following stopping criteria is met: (1) there is no higher C value for a new split; (2) at the first split, the absolute value of the mean difference in treatment outcome is smaller than a pre-specified critical minimum value d_{min}; (3) the pre-set minimal sample size per treatment condition is reached; or (4) the maximum number of leaves is exceeded.

3.2.2 Pruning Algorithm

In the pruning process, the bias-corrected bootstrap is applied to assess the performance of the tree of size T (Dusseldorp and Van Mechelen, 2014). For a T size tree developed at the splitting step, we can obtain the corresponding global partitioning criterion, $C_{T,total}$. When it comes to the bootstrap procedure, a bootstrap sample is drawn and a tree of the same size T is grown. After calculating the corresponding partitioning criterion, $C_{T,boot}$, for the bootstrap sample, we evaluate the performance of the exact bootstrap tree (the same split variable, split point and subgroup assignment) in the total samples. By using the original samples to grow an exact same structure tree, partitioning criterion, $C_{T,orig}$, can be obtained.

For a fixed size tree, the over-optimism value is defined as:

$$O_{b,T} = C_{T,boot} - C_{T,orig}.$$

The over-optimism value shows the difference in the performance values between the bootstrap samples and the original samples. Averaging across all the bootstrap samples $b\,(b = 1, \ldots, B)$, we have the average over-optimism value:

$$\bar{O}_T = \frac{1}{B}\sum_{b=1}^{B} O_{b,T},$$

with a standard error $SE_T = \sqrt{\frac{\sum_{b=1}^{B}(O_{b,T}-\bar{O}_T)^2}{B-1}}$. Subtracting the over-optimism element from the global partitioning criterion $C_{T,total}$ obtained at the first step, we end up with the bias-corrected performance value. It is defined as:

$$C_T^{bc} = C_{T,total} - \bar{O}_T.$$

After calculating the bias-corrected performance values for all tree sizes, $T\,(T = 1, \ldots, T)$, the best subtree is chosen based on 1-SE cross-validation. That is, we first draw a 1-SE range for the highest bias-corrected performance value by cross-validation. Then, the best subtree is the smallest tree in that 1-SE range.

3.3 Subgroup Identification based on Differential Effect Search

Lipkovich et al. (2011) proposed a recursive partitioning method called subgroup identification based on differential effect search (SIDES), which aims at identifying subgroups of subjects with enhanced treatment effect. SIDES directly searches for the subgroup with a

maximum treatment effect, and assigns the rest as the complementary subgroup (Lipkovich et al., 2011). Slightly different from the first two methods, the algorithm of SIDES contains the following three steps: splitting criterion, continuation criterion and selection criterion.

3.3.1 Splitting Criterion

At the beginning of the splitting procedure, we can choose to divide the dataset into one training set and $H-1$ validation sets based on a balanced allocation with respect to the treatment and the covariates, where H is a pre-specified value (Lipkovich et al., 2011). The training dataset is used to identify the potential subgroups and the testing dataset is used to confirm the subgroups.

At the first step, SIDES evaluates all the covariates and their corresponding possible binary splits by following algorithm. For categorical variables, the splits take into account the variable types, such as ordinal or nominal variables. The continuous variables are divided into a pre-defined number of classes. SIDES evaluates the splits based on one of the following splitting criteria (Lipkovich et al., 2011):

$$
\begin{aligned}
p_1 &= 2\left[1-\phi\left(\frac{|Z_{E1}-Z_{E2}|}{\sqrt{2}}\right)\right], \\
p_2 &= 2\min\left(1-\phi(Z_{E1}), 1-\phi(Z_{E2})\right), \\
p_3 &= \max(p_1, p_2),
\end{aligned}
$$

where $\phi()$ is the cumulative distribution function of the standard normal distribution. Z_{E1} and Z_{E2} represent the Z-test statistics for testing one-sided hypothesis for the treatment contrast in the two child nodes, respectively. The first criterion tries to maximize the differential effect between the two child subgroups, and the second one maximizes the treatment effect in one of the two child subgroups. The third criterion is a combination of the first two. There is also a fourth criterion evaluating both the efficacy and the safety. The safety evaluation is performed to search for patients who experience a higher level of adverse effects

or other safety issues. The fourth criterion aims to achieve a balance between the efficacy and the safety, and is defined as:

$$p_4 = 2\left[1 - \phi\left(\frac{|\omega(Z_{E1} - Z_{E2}) + (1-\omega)(Z_{S2} - Z_{S1})|}{\sqrt{2(\omega^2 + (1-\omega)^2)}}\right)\right],$$

where ω is a non-negative weight. Z_{S1} and Z_{S2} are testing the one-sided hypothesis for the safety.

Since different categorical variables have different numbers of possible splits, the variable with the largest number of possible splits has the highest chance to be chosen. This introduces a selection bias into the splitting procedure. The multiplicity adjustment based on the *Šidák* test is applied to control the selection bias, and the adjusted p_a value is defined as (Lipkovich et al., 2011):

$$p_a = 1 - (1-p)^{G^*}$$

where G^* is the effective number of splits taking into account the correlation among the p values for covariates. G^* is defined as $G^* = G^{1-\bar{r}}$, where G is the actual number of splits and $\bar{r}$ is the average correlation among p values from the $G(G-1)/2$ potential splits (Lipkovich et al., 2011).

All pairs of subgroups are ordered from the best to worst in terms of the adjusted p_a values and the best S pairs are retained for the next step. S is a pre-set maximum number of best promising subgroups with a default value of five.

3.3.2 Continuation Criterion

Within the S pairs of promising subgroups, a continuous procedure is applied sequentially, starting with the subgroup with the smallest p_a value (Lipkovich et al., 2011). For the selected subgroup, SIDES considers it the parent node with a treatment effect p_p, and calculates the differential effect for its possible child nodes p_c after excluding the covariate which is already included in the parent node. Once all p_c for the possible child nodes are obtained, the best S

pairs are held out for testing the continuation criterion. The continuation criterion is simply checking if the child node enlarges the differential effect in an expected scale compared to the parent node, by the following definition:

$$p_c \leq \gamma \, p_p,$$

where $0 < \gamma \leq 1$ is a pre-defined relative improvement parameter. A smaller value indicates a stricter selection procedure. If the continuation criterion is met, the child node is added to the parent node and the result from this two-level split is considered as a candidate subgroup.

The splitting procedure is recursively applied to the promising subgroup until one of the pre-defined stopping criteria is met: (1) the number of covariates used to define a subgroup reaches the pre-specified maximum value L; or (2) the subgroup size is smaller than the minimum desired number (Lipkovich et al., 2011). Without further splitting, the current node becomes the terminal node. If the identified subgroup has a p-value smaller than the pre-set significant level α, this subgroup will be included in a candidate subgroup pool. The largest number of candidate subgroups generated from this algorithm can be $S+S^2+\cdots+S^L$ (Lipkovich et al., 2017).

3.3.3 Selection Criterion

After the iteration of the splitting and continuation procedures, we end up with a set of candidate subgroups S_j, where $j = 1, \ldots, J$. The testing dataset is used to confirm these candidates and finalize a number of promising subgroups. If the differential effect, p_t, obtained from the testing dataset is still smaller than the significant level, this subgroup is confirmed as a promising subgroup.

If the testing dataset is not specified, a permutation procedure is applied to verify the subgroups. This procedure aims to control the overall type I error by averaging the results from a large number of permutation datasets. Let p_j denote the original p-value for the

subgroup, S_j. p_k is calculated to evaluate the treatment effect of K permuted datasets, where $k = 1, \ldots, K$. Lipkovich et al. (2017) defined the adjusted p-value as:

$$\tilde{p}_j = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(p_k \leq p_j).$$

Subgroups with $\tilde{p}_j < \alpha$ are identified as promising subgroups. These subgroups represent patients who benefit the most from a certain treatment.

SIDES works well with a relatively small number of candidate covariates (around 15-20), while performance tends to deteriorate for a larger number of covariates. For any study with a large number of covariates, Lipkovich and Dmitrienko (2014b) also proposed an improved two-stage method called SIDEScreen to identify subgroups.

3.4 Handling Missing Data

For the treatment effect methods we introduced, there are default ways of handling missing data within IT and QUINT. IT excludes the covariate containing missing value from the splitting procedure, and QUINT applies listwise deletion which is simply excluding the entire record if there is a single missing value. SIDES does not introduce any procedure to deal with missing values.

We adopt the listwise deletion and three imputation methods to address missing data. Imputation methods are a general way of handling missing data and can be classified as single imputation and multiple imputation methods (Little and Rubin, 2014). Single imputation fills in each missing value with one single prediction while multiple imputation generates M set of predictions.

3.4.1 Single Imputation

For single imputation, we adopt a non-parametric method called MissForest (MF), which is developed from the random forest framework (Stekhoven and Bühlmann, 2011). MF can handle mixed-type, high dimensions, complex interactions and non-linear data structures (Stekhoven and Bühlmann, 2011). This method divides the dataset into two parts, treating the complete records as the training data and the missing values as the testing data. For a data matrix $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ and an arbitrary covariate $\boldsymbol{X}_m$ that includes missing values, MF treats $\boldsymbol{X}_m$ as the response variable, and the other variables $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{m-1},\ \boldsymbol{X}_{m+1}, \ldots, \boldsymbol{X}_p)$ as the predictive variables. $\boldsymbol{X}_m^{(obs)}$ denotes the observed part of the covariate $\boldsymbol{X}_m$, and $(\boldsymbol{X}_1^{(obs)}, \ldots, \boldsymbol{X}_{m-1}^{(obs)}, \boldsymbol{X}_{m+1}^{(obs)}, \ldots, \boldsymbol{X}_p^{(obs)})$ represent the corresponding values for the other variables. Similarly, $\boldsymbol{X}_m^{(mis)}$ represents the missing part of the $\boldsymbol{X}_m$, while $(\boldsymbol{X}_1^{(mis)}, \ldots, \boldsymbol{X}_{m-1}^{(mis)}, \boldsymbol{X}_{m+1}^{(mis)}, \ldots, \boldsymbol{X}_p^{(mis)})$ are the corresponding values for the same observation. At the first step, MF fits the training data to a random forest $\boldsymbol{X}_m^{(obs)} \sim \boldsymbol{X}_1^{(obs)}, \ldots, \boldsymbol{X}_{m-1}^{(obs)},\ \boldsymbol{X}_{m+1}^{(obs)}, \ldots, \boldsymbol{X}_p^{(obs)}$; then predicts the missing values $\boldsymbol{X}_m^{(mis)}$ by applying the trained random forest to the $\boldsymbol{X}_1^{(mis)}, \ldots, \boldsymbol{X}_{m-1}^{(mis)},\ \boldsymbol{X}_{m+1}^{(mis)}, \ldots, \boldsymbol{X}_p^{(mis)}$. MF runs the above algorithm repeatedly and stops if the difference between the newly imputed data and the previous one within a threshold.

3.4.2 Multiple Imputation

The obvious disadvantage of single imputation is that imputing a single value cannot reflect the sample variability for missingness. Multiple imputation (MI) methods address the uncertainty by replacing each missing value with M plausible values, resulting in M complete data sets (Little and Rubin, 2014). Each complete data set is analysed by the same complete-data method and the results are pooled to obtain a combined estimate. We introduce two MI methods into our study: MI based on Predictive Mean Matching (MIPMM) and MI based on Expectation-Maximization (MIEM). We use $M = 5$ for both MIPMM and MIEM approaches.

The MIPMM method is a semi-parametric imputation approach which defines the pre-

dictive mean of an observation as $\hat{Y}_i = X_i^T \beta$. At first, this method calculates the predictive mean for all observations. Then, when there is an incomplete case, MIPMM randomly draws an observation with a similar predictive mean from the complete cases to substitute for the missing one (Schenker and Taylor, 1996).

The MIEM method assumes that the data follows a multivariate normal distribution $N(\mu, \Sigma)$ and the missing values are consistent with the MAR mechanism. This method combines the classic EM algorithm with a bootstrap approach (Ghorbani and Desmarais, 2017). The classic EM algorithm is an iterative procedure requiring two steps at each iteration: an expectation step (E-step) and a maximization step (M-step) (Levine and Casella, 2001). The E-step is a calculation of a conditional expectation and the M-step is a maximization of this expectation with respect to relevant parameters. For the MIEM method, it bootstraps M datasets, and then runs the EM algorithm to predict point estimates of μ and Σ for each bootstrap dataset (Honaker and King, 2010; Honaker et al., 2011). Once it converges, MIEM imputes the missing values based on its distribution conditional on the original observed values and each set of estimates (Honaker et al., 2011).

After using these methods to address missing data, the performance of the three treatment effect methods is evaluated in simulation studies.

Chapter 4

Simulation Study

Simulation studies are used to compare the model performance in the treatment effect when using complete, incomplete and imputed data across various scenarios when the sample size, proportion of missingness and missing variables differ. This chapter covers the design and results of the simulation studies.

4.1 Study Design

Based on different sample sizes (100, 500), proportions of missingness (0.2, 0.5) and missing variables (missing only in X_1, or missing in X_1 and X_2), we considered eight scenarios in our simulation study (Table 4.1). For each scenario, we generated two complete datasets consisting of n ($n = 100$, or 500) observations on a binary treatment (T), and ten continuous variables ($X_1, \ldots, X_{10}$) simulated from a multivariate normal distribution. T is equal to 0 for the control treatment group and 1 for the experimental treatment group. Consider μ_{x_p} and μ_Y as the means for the predictor variables and the outcome variable, and Σ_X and σ_Y as the variance-covariance matrix and the variance, respectively. μ_{x_p} for the splitting variables ($\mu_{x_1}, \mu_{x_2}, \mu_{x_3}$) is fixed at 5, 10, and 1, respectively. μ_{x_p} for the other variables are drawn from different uniform distributions, which are: $\mu_{x_5} \sim U(-5, 5)$, μ_{x_4}, μ_{x_6}, μ_{x_7} and μ_{x_9} are generated from $U(-10, 10)$, and μ_{x_8} and $\mu_{x_{10}}$ followed a uniform distribution on the range

of $(-1, 1)$. Σ_X is a 10×10 matrix with the main diagonal equal to 1 and the rest of the entries equal to 0.5:

$$\Sigma_X = \begin{bmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \cdots & 1 \end{bmatrix}.$$

Table 4.1: Detailed Information for Eight Scenarios

Scenario	Sample Size	Proportion of Missingness	Missing Variable
1	500	0.2	X_1
2	500	0.5	X_1
3	100	0.2	X_1
4	100	0.5	X_1
5	500	0.2	X_1 and X_2
6	500	0.5	X_1 and X_2
7	100	0.2	X_1 and X_2
8	100	0.5	X_1 and X_2

For the first model, the outcome Y is a continuous variable generated from a normal distribution with $\sigma_Y = 1$ and μ_Y depending on each leaf and the treatment condition (M1, Figure 4.1). In this model, d represents the Cohen's effect size which indicates the standardized difference between the two means for different treatments (Dusseldorp and Van Mechelen, 2014). This value d is a perfect indication for the heterogeneity of treatment effects across subgroups. For example, in Figure 4.1, the first two groups show a qualitative treatment interaction effect, since d in the first group is positive and in the second group it is negative. The third and the fourth groups contain a quantitative interaction, since the d values only differ in the magnitude.

The second model (M2) is generated from a linear regression:

$$Y = 2 + 2 \cdot T + 2 \cdot x_1 - 2 \cdot x_2 + 2 \cdot x_3 + 3 \cdot T \cdot z_1 - 3 \cdot T \cdot z_2 + 3 \cdot T \cdot z_3 + \epsilon$$

where $\epsilon \sim N(0, 1)$. In order to be comparable to the first model, indicator variables (z_1, z_2, z_3), where $z_1 = \mathbb{1}\{x_1 > 5\}$, $z_2 = \mathbb{1}\{x_2 > 10\}$, $z_3 = \mathbb{1}\{x_3 > 1.3\}$, are introduced into the regression model to include the same splitting point information in the treatment-covariate interaction. The results are compared to the ones obtained from the first model, to see if the data structure influences the performance of the treatment effect models and the data completion methods.

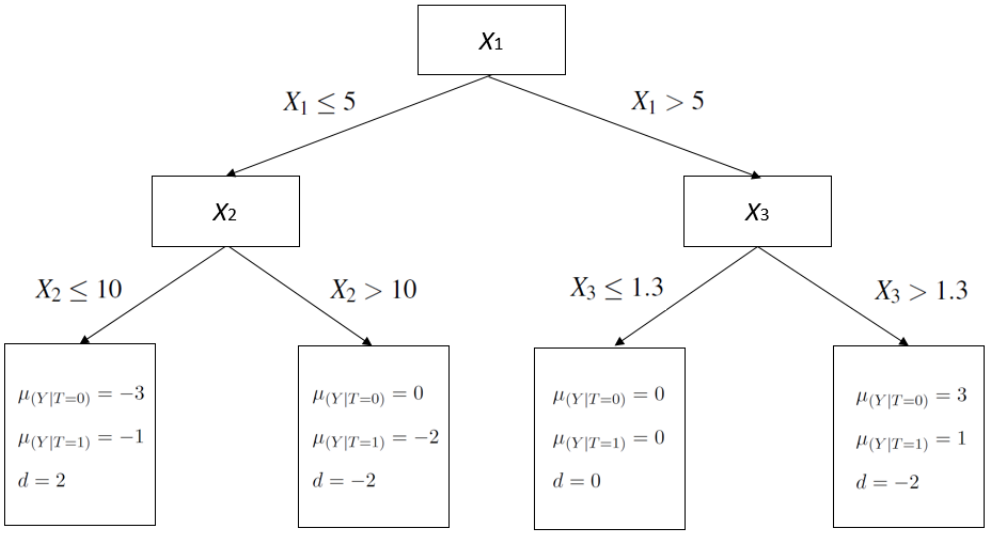

Figure 4.1: Model 1: Tree Structure with Continuous Outcomes

4.2 Model Performance

We used two methods to compare the model performance between the complete, incomplete and imputed datasets in each scenario. The first one estimates the accuracy of the variable selection (Loh et al., 2015). Since our true splitting variables are X_1, X_2 and X_3, we compare the frequencies of those variables chosen at the root node and two child nodes. We also

calculate the bias in the treatment effect to compare these three approaches under different models (Loh et al., 2016). Given a node j in a tree T, let $\mu(z, j)$ denote the mean of Y for the treatment z in node j, and $\hat{\mu}(z, j)$ be its estimate. $\tilde{T}$ and $|\tilde{T}|$ represent the set of terminal nodes of T and their corresponding number, respectively. We evaluated the model performances based on the average error and average relative error as defined in Loh et al. (2016):

Average error:

$$D(T, z) = |\tilde{T}|^{-1} \sum_{j \in \tilde{T}} \big[\hat{\mu}(z, j) - \mu(z, j)\big],$$

Average relative error:

$$R(T, z) = |\tilde{T}|^{-1} \sum_{j \in \tilde{T}} \big[\hat{\mu}(z, j) - \mu(z, j)\big] / \mu(z, j).$$

For our simulation study, we averaged the results from 100 simulations, which gave us the average bias and average relative bias (Loh et al., 2016).

4.3 Results

As mentioned in the previous section, we averaged the results from 100 simulations to compare model performance. This section shows the model performance results for eight different scenarios with varied samples sizes, proportions of missingness, and missing variables. For each scenario, we evaluated the frequencies of variable selection and the average bias/relative bias.

4.3.1 Scenario One

The first scenario has a relatively large sample size with 500 observations. We generated MAR missing values for variable X_1 which depends on X_3, with the proportion of missingness equal to 0.2. For the MAR values, a probability distribution for missingness over X_3

was generated first, and then the corresponding values with a probability smaller than 0.2 were excluded from X_1. Table 4.2 shows the performance matrix of the two models when using different imputation methods addressing missing data, and Figures 4.2 and 4.3 are the corresponding plots. Since SIDES only successfully detected subgroups from one simulation out of 100 for M1 after using the MissForest method to deal with the missing data, we believed that the results were not comparable, so we removed them from the plots. The complete plot including the results is shown in the appendix Figure A.1.

The variable selection frequency shows that all three methods perform equally well for the first model at the first split, with 100% accuracy rates for all imputation methods, while the results of SIDES start falling behind after the second split. For the second model, SIDES and IT outperform the QUINT dramatically at the first split, with the accuracy rates increased by 70%. Unfortunately, SIDES only works when using the MissForest approach to address missing data. Furthermore, the performance of QUINT is improved slightly when the MissForest method is used to handle missing data. Similar to the first model, SIDES is not good at detecting the second and the third split. In contrast, the QUINT starts working well at the last two splits and shows an almost equivalent performance with IT.

When it comes to bias, we expected a smaller bias and relative bias when using the imputation methods to address the missing data. However, QUINT has a large relative bias for the first model, especially for the the control treatment group when using the datasets imputed by the MIEM and MIPMM approaches (with relative biases equal to 2.8 and 2.0, respectively). For the second model, QUINT in the experimental treatment group has a large bias and relative bias. Both of them are larger when the dataset is imputed by the MissForest method. IT method also shows slightly large relative biases for the first model, but these biases are smaller than 1 and are acceptable.

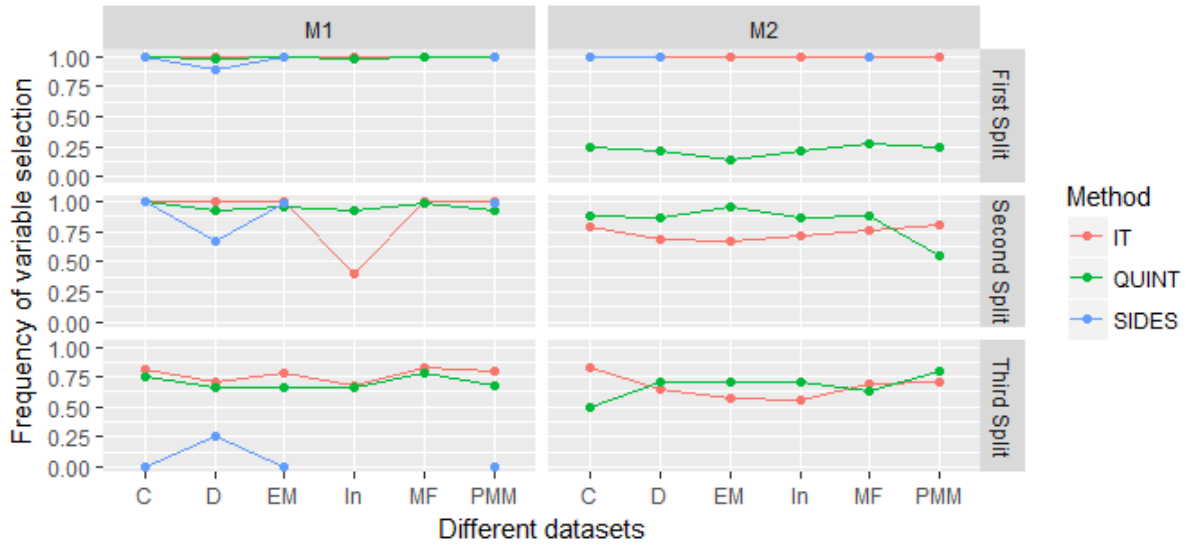

Figure 4.2: The Frequency of Variable Selection for the First Scenario [2]

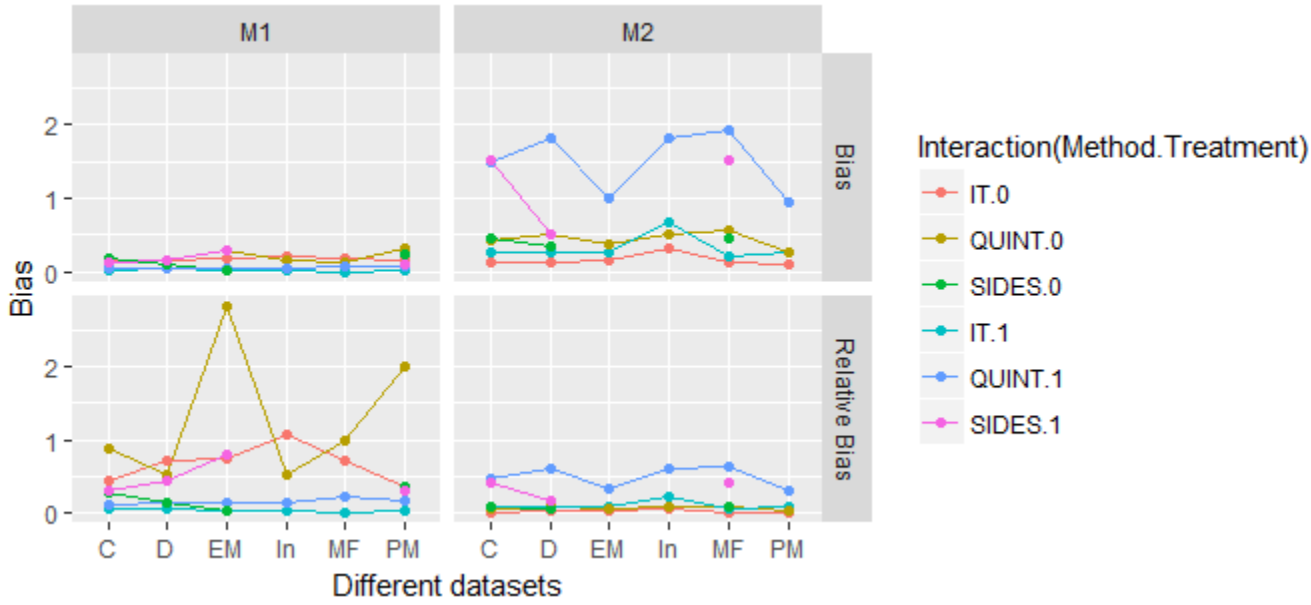

Figure 4.3: The Bias and Relative Bias for the First Scenario [3]

[2]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[3]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.3.2 Scenario Two

The second scenario has the same sample size and missing variable as the first scenario, which are 500 samples and MAR values for variable X_1 which depends on X_3. We generated a larger proportion of missingness in this scenario, which was 0.5. Table 4.3 shows the performance matrix of the two models when using different imputation methods addressing missing data. Figures 4.4 and 4.5 are the corresponding plots. Similar to the first scenario, we removed the results from the SIDES for M1 after using the MissForest method. The complete plot with the results is shown in the appendix Figure A.2.

The result from the SIDES method shows similar trends to the one derived from the first scenario. This method only works well at the first split for the first model, with 100% accuracy rates. For the second model, SIDES only works when using the MissForest to address the missing data. Similar to the first scenario, IT performs slightly better than QUINT at the first split of the tree-based model (M1), but performs dramatically better when using the regression-based model (M2), with more than 50% increased in the accuracy rates. However, the QUINT method starts working well at the second split and even performs slightly better than IT at some points. The results from the first two scenarios show that, with a relatively large sample size, the proportion of missingness does not substantially influence the model performance.

For the bias, similar to the first scenario, QUINT has the highest average relative bias for the first model, especially for the datasets imputed by the MIEM and MIPMM approaches, with both relative biases larger than 2. QUINT also shows a highest bias and relative bias in the second model for the experimental treatment group, regardless of the imputation methods used. Similar to the first scenario, IT method shows slightly large relative biases for the first model.

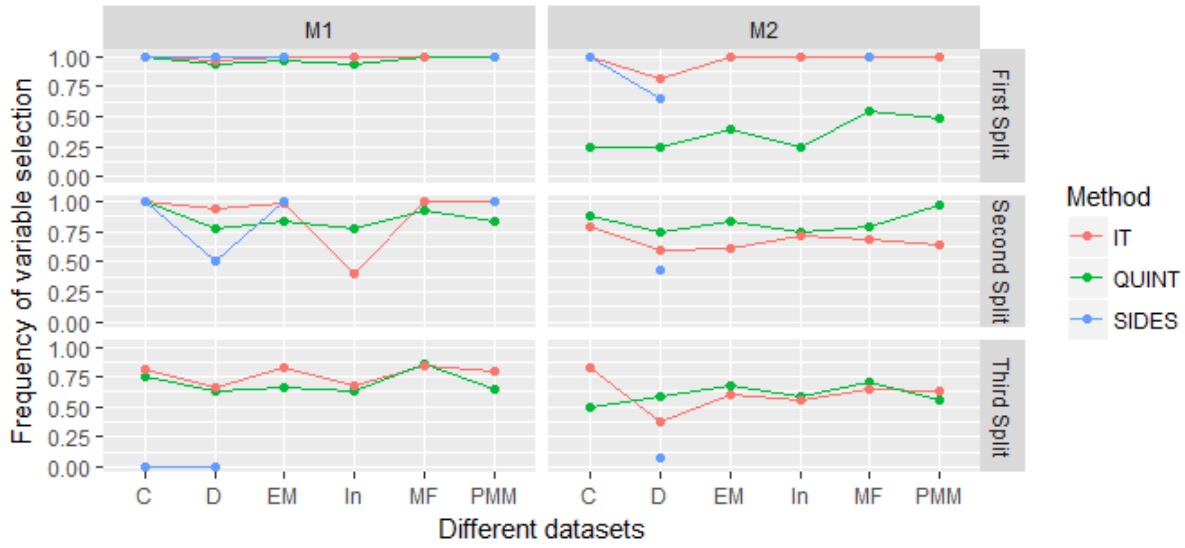

Figure 4.4: The Frequency of Variable Selection for Scenario Two [5]

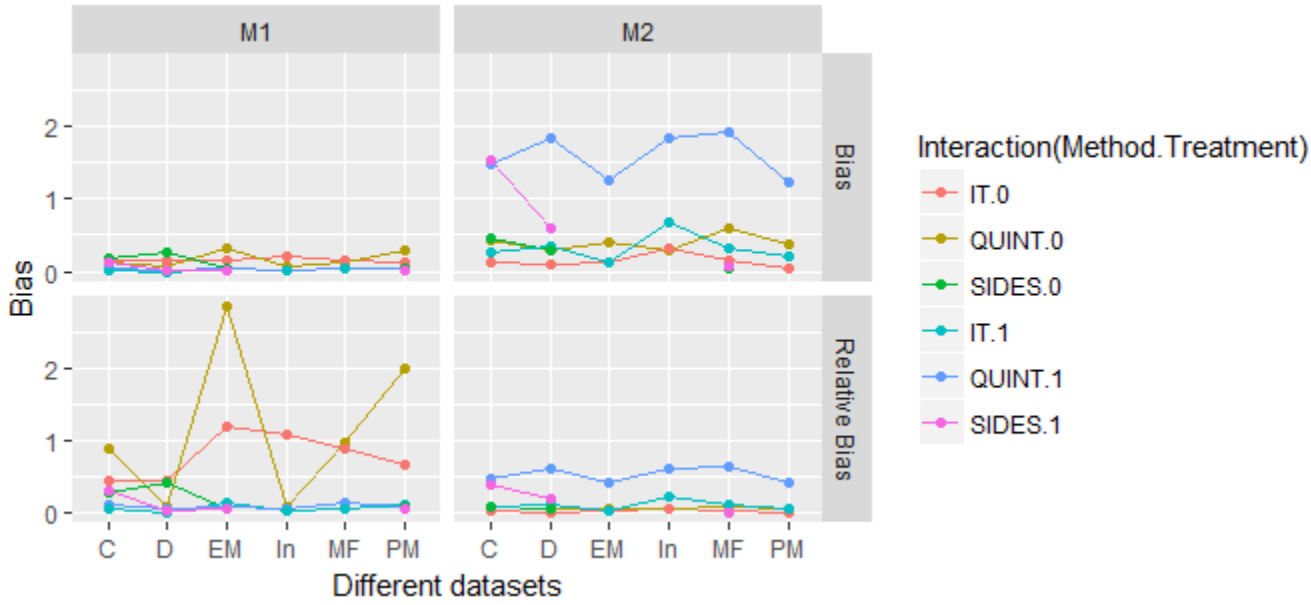

Figure 4.5: The Bias and Relative Bias for Scenario Two [6]

[5] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[6] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.3.3 Scenario Three

The third scenario has a smaller sample size compared to the first two scenarios, with only 100 observations. We generated MAR missing values for variable X_1 which depends on X_3, with the proportion of missingness equal to 0.2. Table 4.4 shows the performance matrix of the two models when using different imputation methods addressing the missing data. Figures 4.6 and 4.7 are the corresponding plots.

A smaller sample size identified a decrease in model performances, especially for the IT and QUINT methods, with more than 10% decreased in accuracy rates across all splits. However, the conclusions are similar to the ones drawn from the first two scenarios. First, after using the imputation methods to address the missing data, the SIDES method performs better than the IT and QUINT methods when using the tree-structure data (M1). For the regression-based data (M2), SIDES only works well when using the MissForest method to address the missing data. Second, IT outperforms the QUINT method at the first split as before, while QUINT narrows down the performance gap at the second split.

There are similar conclusions for the bias for scenario three: QUINT has the highest relative bias for the first model, especially for the control treatment group when using the listwise deletion to address the missing data. Compared to the previous scenarios, the relative bias for IT remains the same, which is close to 1 when using the dataset imputed by the MIEM method. For the second model, QUINT has a large value for the bias and relative bias for the experimental treatment group.

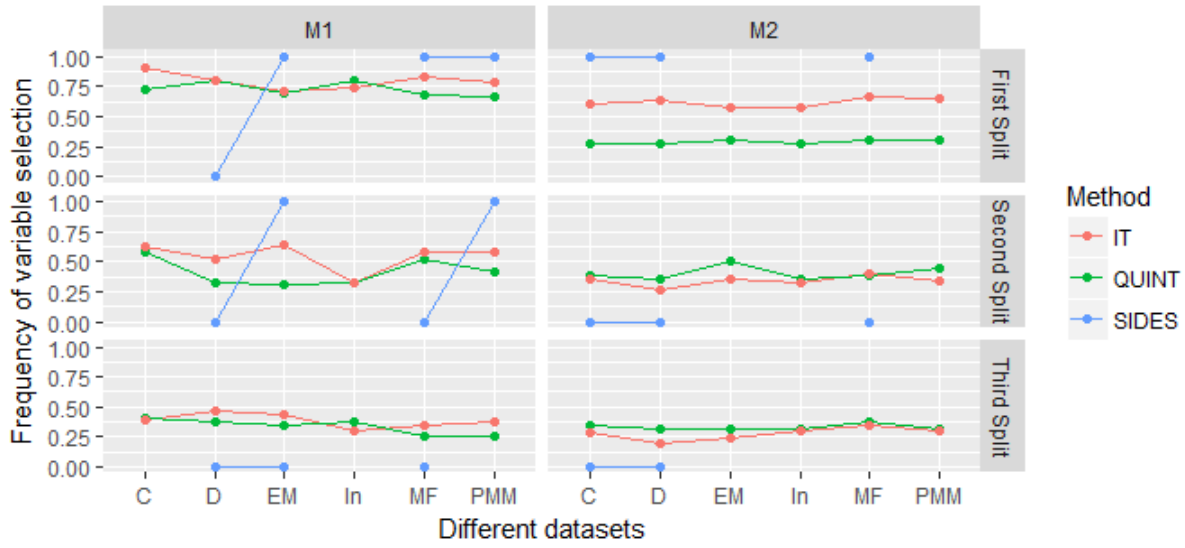

Figure 4.6: The Frequency of Variable Selection for Scenario Three [7]

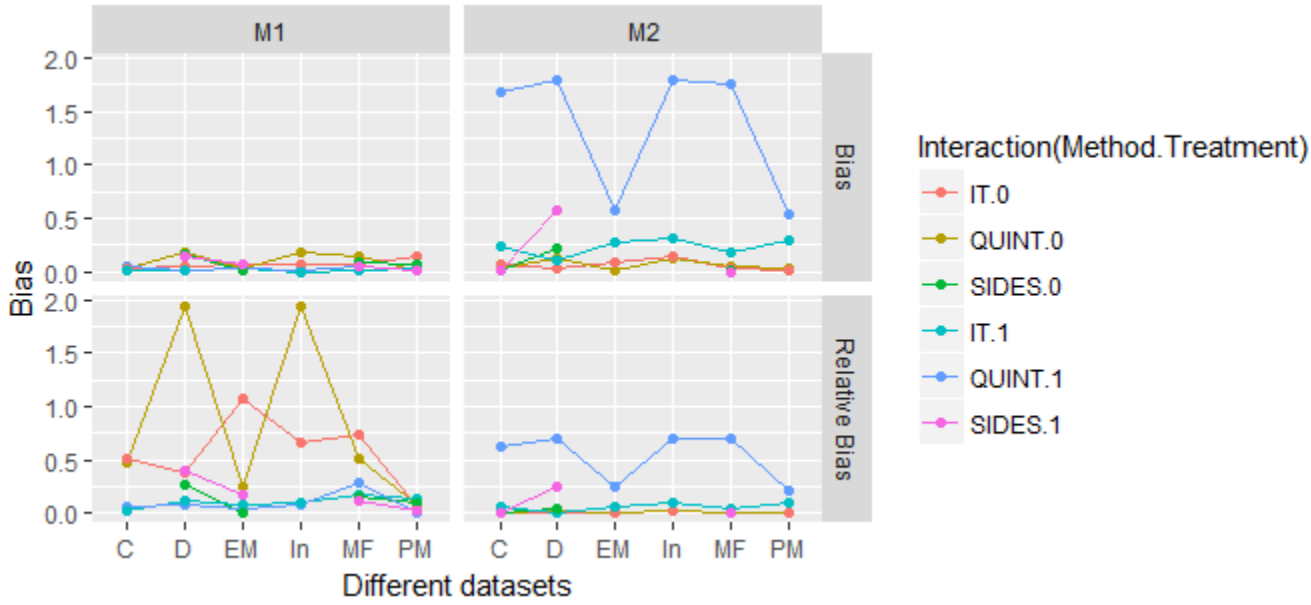

Figure 4.7: The Bias and Relative Bias for Scenario Three [8]

[7] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[8] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.3.4 Scenario Four

The fourth scenario has the same sample size and missing variable as the third one, with 100 samples and MAR for variable X_1. The proportion of missingness is now increased to 0.5. Table 4.5 shows the performance matrix of the two models when using different imputation methods to address the missing data. Figures 4.8 and 4.9 are the corresponding plots. After using listwise deletion to address missing values, SIDES only successfully detected subgroups from one simulation run for M1. We believed that the results were not comparable to the other methods, so we removed these results from the plots. The complete plot with the results is shown in the appendix Figure A.3.

In a scenario with a smaller sample size and a larger proportion of missingness, SIDES still performs better than IT and QUINT at the first split for the tree-based model (M1). However, for the regression-based model (M2), the SIDES method only works well when using MissForest to deal with the missing data. The IT performs better than QUINT at the first split, especially for the second model, with a 30% higher in the accuracy rate, while QUINT starts working well at the last two splits, and even performs better than IT from time to time.

The results for the biases are consistent with the ones obtained from the third scenario, which indicates that the biases are more dependent on the sample size rather than the proportion of the missingness.

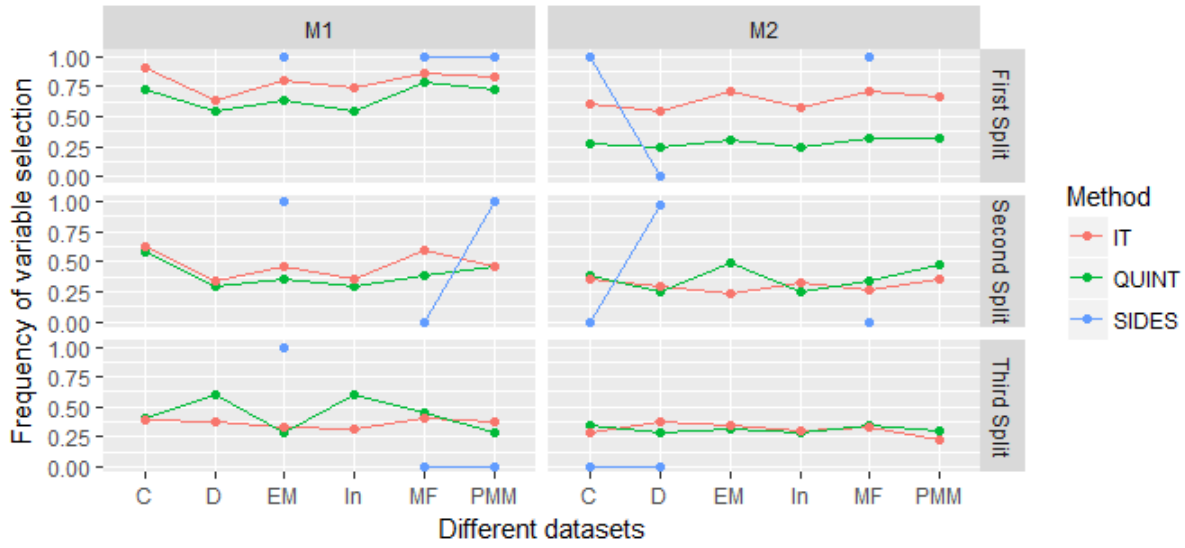

Figure 4.8: The Frequency of Variable Selection for Scenario Four [10]

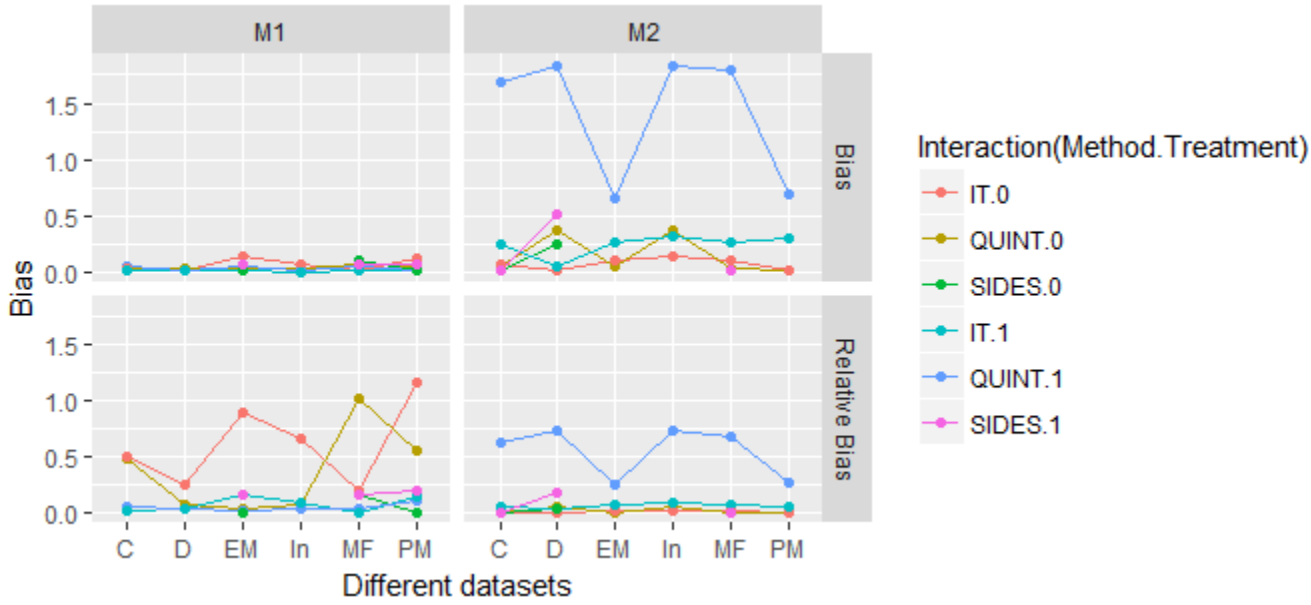

Figure 4.9: The Bias and Relative Bias for Scenario Four [11]

[10] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[11] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.3.5 Scenario five

In the fifth scenario, we increased the number of missing variables. We generated MAR for variables X_1 and X_2, which depend on X_3 and X_4, respectively. For the MAR values, probability distributions for missingness over X_3 and X_4 was generated respectively, and then the values with a probability smaller than 0.2 were excluded from X_1 and X_2 correspondingly. Both proportions of missingness are equal to 0.2, and the sample size is 500. Table 4.6 shows the performance matrix of the two models when using different imputation methods addressing missing data. Figures 4.10 and 4.11 are the corresponding plots.

Doubling the number of missing variables show no impact on model performance. Similar to the first scenario, all three methods perform equally well at the first two splits for the first model, while the results from the SIDES method start to fall behind at the third split. The second model shows that IT outperforms QUINT at the first split, while QUINT starts to narrow down the performance gap from the second split and shows a similar performance with IT from time to time. However, SIDES fails to detect any subgroup in the regression-based model.

For the biases, the results are consistent with the previous scenarios. There is no observed difference in bias for IT and SIDES among datasets imputed by different methods.

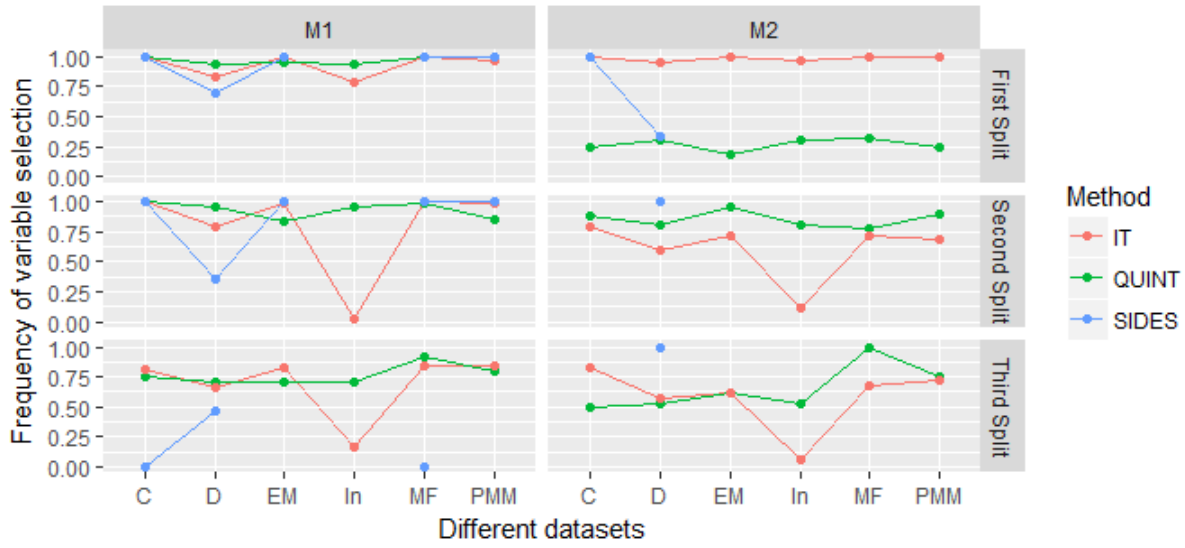

Figure 4.10: The Frequency of Variable Selection for Scenario Five [12]

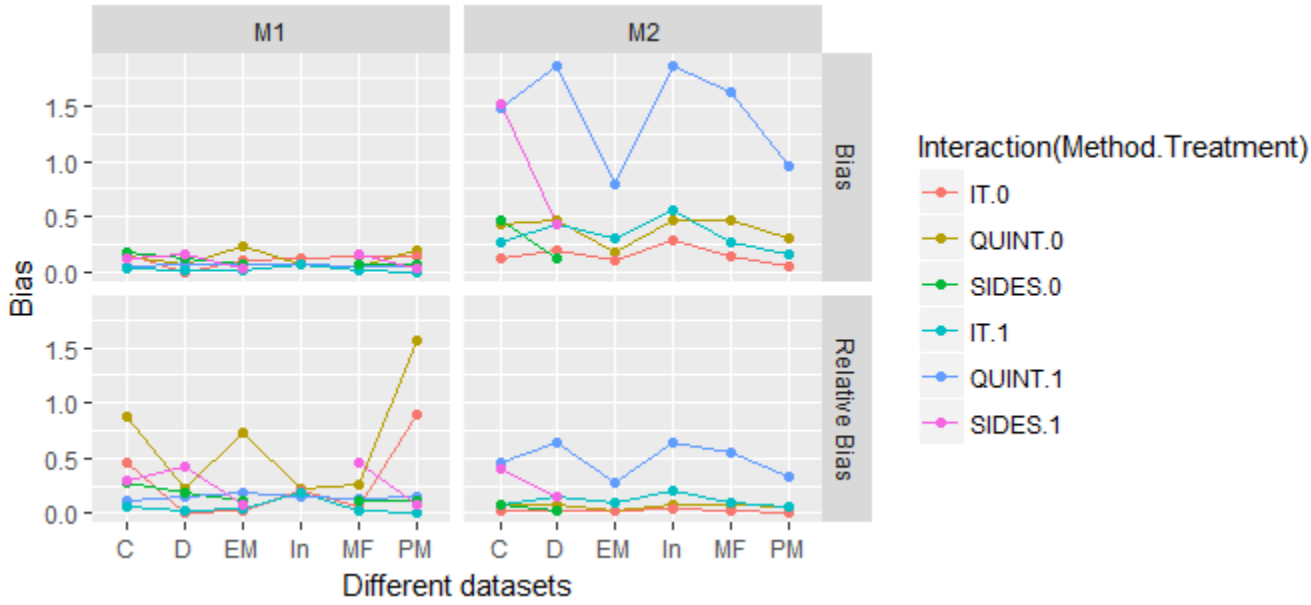

Figure 4.11: The Bias and Relative Bias for Scenario Five [13]

[12] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[13] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.3.6 Scenario Six

The same as in scenario five, we generated MAR for variables X_1 and X_2, which depend on X_3 and X_4, respectively. This scenario has a sample size equal to 500, and a larger proportion of missingness for both missing variables, which are $p = 0.5$. Table 4.7 shows the performance matrix of the two models when using different imputation methods to address the missing data. Figures 4.12 and 4.13 are the corresponding plots.

Similar conclusions can be drawn from this scenario, except for one difference between the first four scenarios and the last two. The SIDES method fails to detect any subgroup from the regression-based model for these scenarios with two variables containing missing data.

Consistent with the previous scenarios, QUINT has the largest relative bias for the first model, especially when using the MIPMM method, with a relative bias up to 1.3. QUINT also shows a large bias and relative bias for the second model, regardless of the imputation methods.

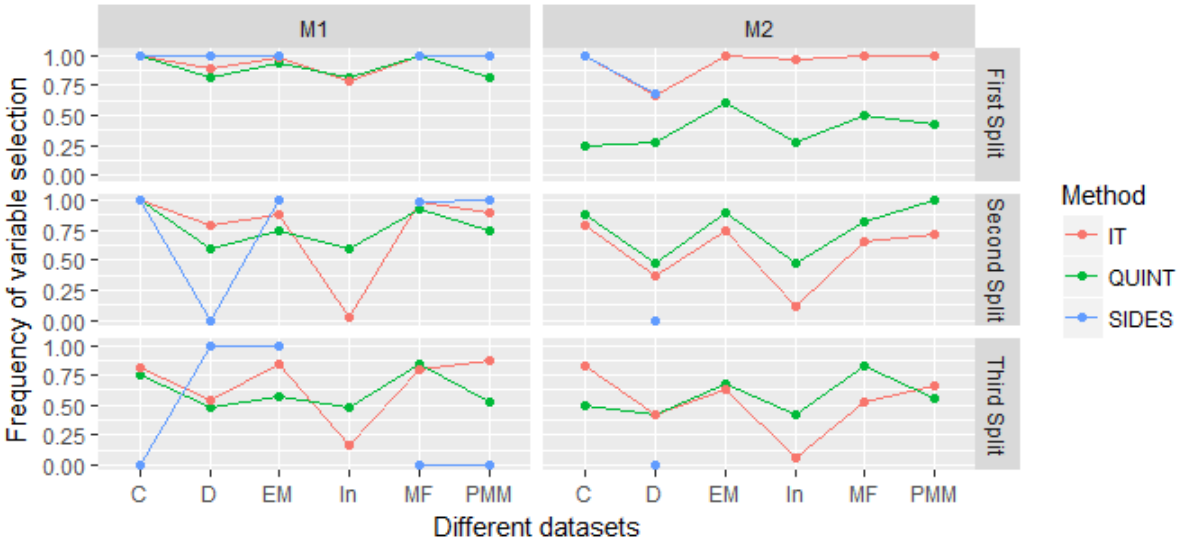

Figure 4.12: The Frequency of Variable Selection for Scenario Six [14]

[14]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

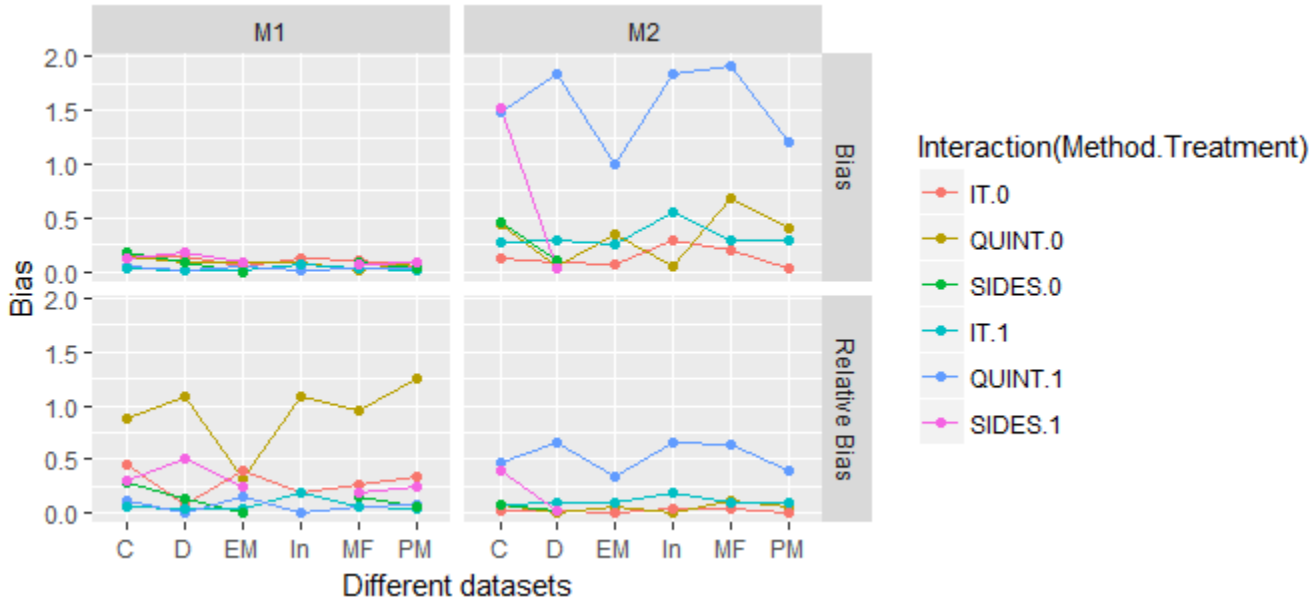

Figure 4.13: The Bias and Relative Bias for Scenario Six [15]

4.3.7 Scenario Seven

We used $n = 100$ as the sample size for scenario seven, and generated MAR for variables X_1 and X_2, which depend on X_3 and X_4, respectively. Both variables' proportions of missingness are equal to 0.2. Table 4.8 shows the performance matrix of the two models when using different imputation methods to address missing data. Figures 4.14 and 4.15 are the corresponding plots. For the second model with the dataset imputed by MissForest, SIDES only detected subgroups from a single simulation. We believed that the results were not comparable to the other methods, so we removed them from the plots. The complete plot with the results is shown in the appendix Figure A.4.

The first split shows that the performance of SIDES is more than 15% higher than the performances of IT and QUINT for the tree-based model. However, SIDES again fails to detect any subgroup for the regression-based model. IT performs better than QUINT at the first split, while QUINT starts working well for the last two splits and performs better than

[15] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

IT occasionally.

When it comes to the bias, the relative biases for both IT and QUINT are increased in the first model, and the QUINT has relatively large bias and relative bias in the second model when using the MissForest to address the missing data, with biases up to 1.8 and 0.7 respectively.

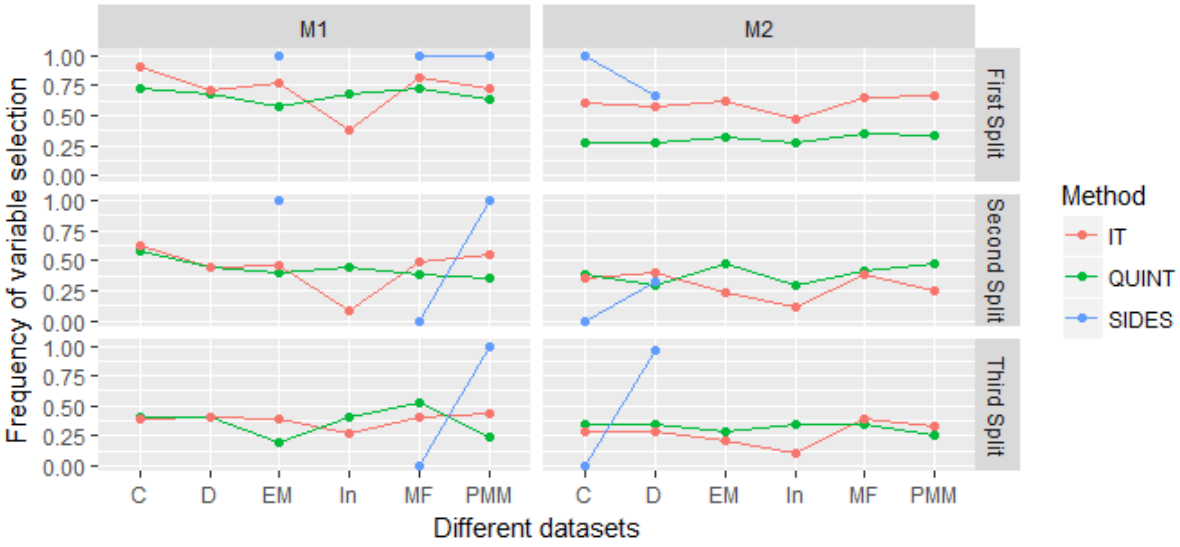

Figure 4.14: The Frequency of Variable Selection for Scenario Seven [17]

[17]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

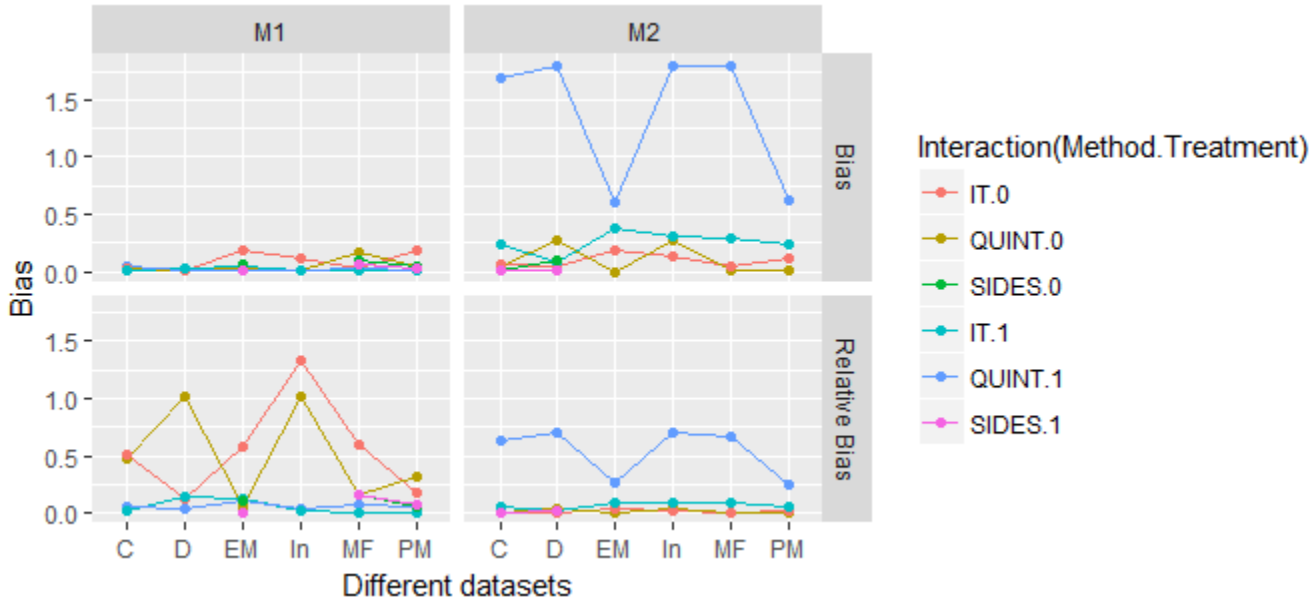

Figure 4.15: The Bias and Relative Bias for Scenario Seven [18]

4.3.8 Scenario Eight

For the last scenario, we increased the proportion of missingness to 0.5. The sample size is 100, and the missing variables are X_1 and X_2 which depend on X_3 and X_4, respectively. Table 4.9 shows the performance matrix of the two models when using different imputation methods to address missing data. Figures 4.16 and 4.17 are the corresponding plots. For the second model with the dataset imputed by MissForest, SIDES only detected subgroups from one single simulation. We believed that the results were not comparable, so we removed them from the plots. The complete plot with the results is shown in the appendix Figure A.5.

For the frequency of the variable selection, the SIDES fails to find any subgroup in most situations, while the other conclusions remain the same as for the previous scenarios. The relative biases for IT for the first model are increased dramatically and are highest among biases for all three methods, especially for the dataset used with the MIPMM approach. The

[18] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

biases for the QUINT and SIDES are the same as for the previous scenarios as well.

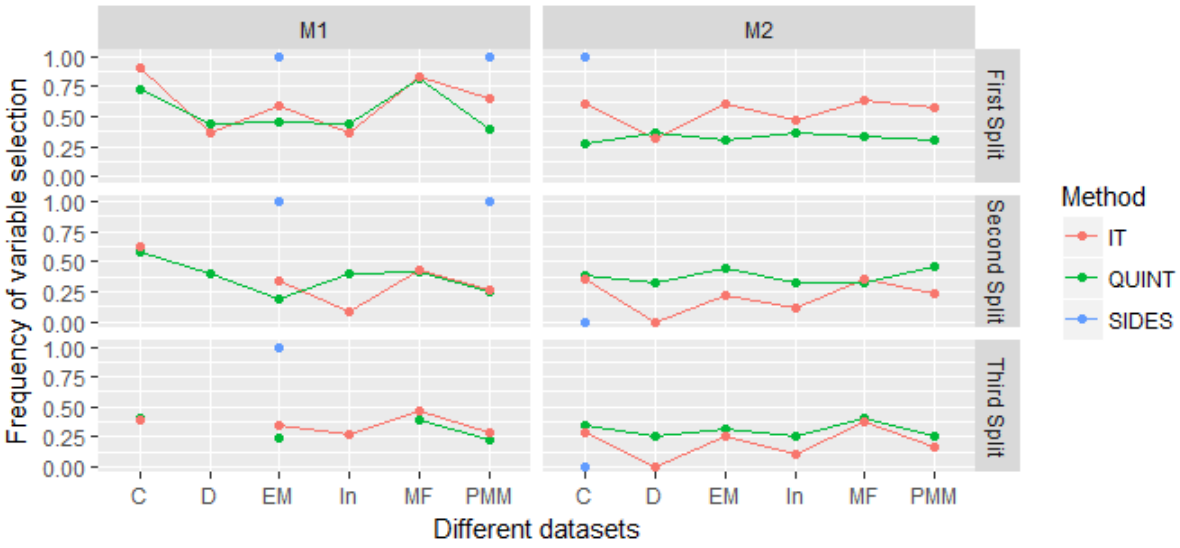

Figure 4.16: The Frequency of Variable Selection for Scenario Eight [20]

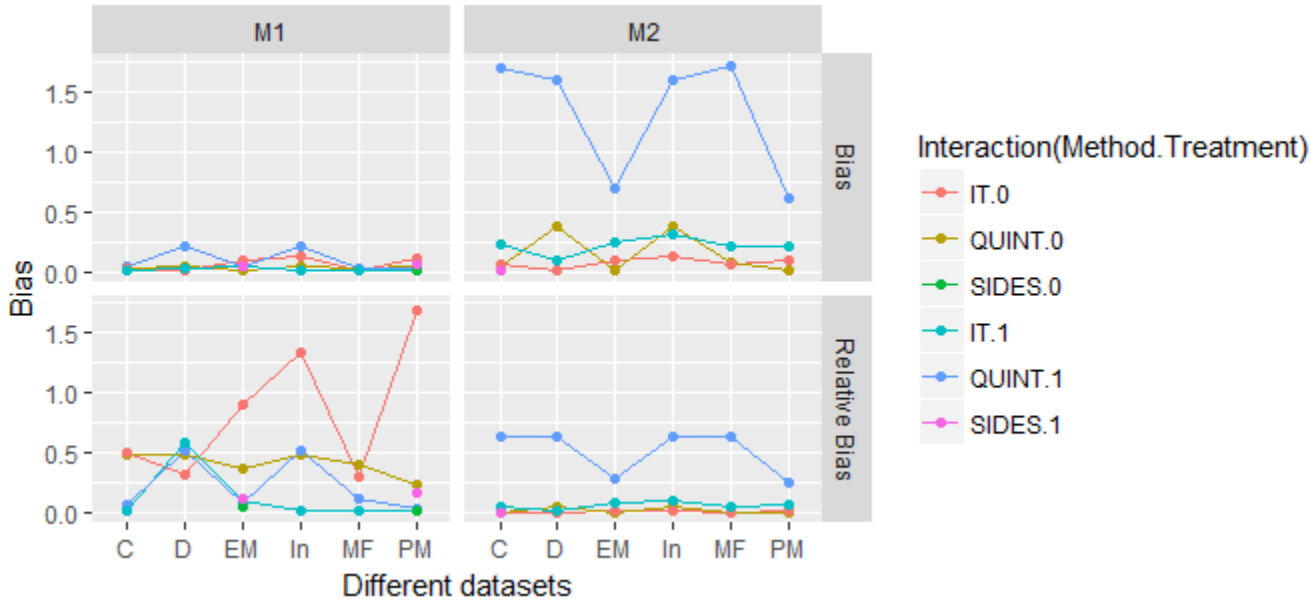

Figure 4.17: The Bias and Relative Bias for Scenario Eight [21]

[20] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[21] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

4.4 Conclusions

For the accuracy of variable selection, SIDES performs better than the IT and QUINT methods at the first split for the model generated from the tree structure, with more than 90% accuracy rates for all eight scenarios. However, for the model generated from the linear regression, the performance of SIDES is satisfactory when using the MissForest to impute the missing data, and the good performance is restricted to the situations with a large number of observations (n=500) and only when one variable contains missing values. The disadvantage of the SIDES method is that it is not sensitive at the second and the third splits, especially for datasets generated from a linear regression method.

The performance of the IT and QUINT methods are approximately equivalent amongst all eight scenarios for the first model, regardless of the sample size, the proportion of missingness and the number of variables with missing data. When it comes to the second model, IT performs dramatically better than QUINT at the first split. However, from the second split, the QUINT method narrows down the performance gap between the two methods and becomes even better than IT at some points. Taken together, the IT and QUINT methods show similar performance and can be used interchangeably for datasets generated from a tree-based structure with a continuous outcome variable. For datasets generated from a linear regression model, IT would be a better choice compared to the QUINT method.

If we keep the other conditions, such as scenarios, models and the level of the split, unchanged, both IT and QUINT perform better when using MissForest to address missing data, especially for the first model. The performances of the two MI methods (MIPMM and MIEM) are equivalent, and the listwise deletion approach would be the last choice for addressing missing data. The SIDES method also shows a similar trend for the first split in both models: the accuracy of variable selection for SIDES is higher when the dataset is imputed by MissForest, regardless of the sample size and the proportion of missingness. The unusual situation that the single imputation method outperforms the MI methods might be due to the data generation approach. We generated the first model from a tree-based

structure, which shares a similar underlying methodology with the MissForest. It could also explain the decrease in the performance in the second model, since we generated the data from a linear regression method instead of a tree structure.

When it comes to the bias for the treatment effect, QUINT has the largest relative bias for the first model, especially for the datasets imputed by the MIEM and MIPMM methods. In addition to that, QUINT with the experimental treatment group has a large bias and relative bias for the second model. The relative bias for IT in the first model begins to increase when the sample size is decreased and the proportion of missingness is increased. Generally speaking, QUINT has a relatively large bias for the datasets imputed by the MI methods, while the other biases are acceptable. It indicates that we need to make a trade-off between the performance and the large biases when applying the QUINT method.

In real life applications, such as clinical trials data with missing values, the treatment effect models should be chosen based on the objective of the study, the sample size, the number of variables containing missing data, and the data structure. For a study that focuses on searching for one particular subgroup who has a better treatment benefit, with a reasonable sample size and only one variable containing missing values, SIDES would be the first option. On the other hand, either IT or QUINT would be a good choice for a study looking for some exclusive subgroups with a tree-based structure. For the rest of the situations, IT would be the best choice, given the consistent performance and the acceptable bias for most cases. For the imputation methods, we should take into account the assumption of the data structure. MissForest is an excellent choice for a dataset with a tree-based structure, while MI methods would be a good fit for the other situations.

Table 4.2: Performance Comparisons for Scenario One Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES [1]
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	100%	100%	100%	98%	99%		98%	100%	89%	99%	100%	99%	99%	99%	99%	99%	99%	100%
	Second Split	100%	100%	100%	93%	40%		93%	100%	67%	92%	100%	99%	96%	100%	99%	99%	100%	100%
	Third Split	75%	81%	0%	66%	68%		66%	71%	25%	67%	79%	0%	66%	78%	0%	78%	83%	0%
	First Split on X_1	0%	11%	100%	0%	0%		0%	16%	89%	0%	13%	99%	1%	8%	99%	0%	15%	100%
Average Bias	$T = 0$	0.137	0.155	-0.178	0.145	0.219		0.145	0.155	-0.093	0.319	0.166	-0.233	0.282	0.173	-0.019	0.132	0.175	-0.132
	$T = 1$	0.052	0.023	0.115	0.056	0.012		0.056	0.034	0.166	0.065	0.018	0.113	0.054	0.012	0.296	0.086	0.003	0.129
Average Relative Bias	$T = 0$	-0.879	-0.456	-0.279	-0.520	-1.068		-0.520	-0.715	-0.146	-2.003	-0.368	-0.363	-2.816	-0.736	-0.030	-0.985	-0.723	-0.212
	$T = 1$	-0.115	-0.055	0.300	-0.137	-0.023		-0.137	-0.072	0.458	-0.159	-0.040	0.301	-0.141	-0.034	0.791	-0.214	0.010	0.348
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	25%	100%	100%	21%	99%		21%	99%	100%	24%	100%		14%	100%		27%	100%	100%
	Second Split	88%	79%		87%	72%		87%	68%		55%	80%		95%	67%		88%	76%	
	Third Split	50%	83%		71%	55%		71%	64%		80%	70%		70%	57%		63%	69%	
Average Bias	$T = 0$	-0.437	0.115	0.466	-0.497	0.329		-0.497	0.137	-0.356	-0.270	0.095		-0.369	0.145		-0.574	0.115	0.466
	$T = 1$	1.482	0.265	1.524	-1.810	0.676		-1.810	0.273	-0.502	-0.945	0.275		-1.008	0.271		-1.926	0.222	1.524
Average Relative Bias	$T = 0$	0.075	-0.019	-0.082	0.084	-0.055		0.084	-0.023	0.059	0.045	-0.016		0.063	-0.024		0.097	-0.019	-0.082
	$T = 1$	0.468	-0.087	-0.404	0.601	-0.237		0.601	-0.089	0.170	0.313	-0.085		0.328	-0.091		0.646	-0.075	-0.404

[1] For the first model with the dataset imputed by the MissForest method, the results of SIDES are not comparable since it only successfully detects subgroups from one single simulation.

Table 4.3: Performance Comparisons for Scenario Two Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES [4]
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	100%	100%	100%	93%	99%		93%	97%	100 %	99%	100%	99%	97%	100%	99%	100%	100%	100%
	Second Split	100%	100%	100%	78%	40%		78%	94%	51%	83%	100%	100%	83%	99%	100%	93%	100%	100%
	Third Split	75%	81%	0%	63%	68%		63%	66%	0%	64%	80%		66%	83%		86%	84%	0%
	First Split on X_1	0%	11%	100%	2%	0%		2%	14%	2%	1%	11%	99%	2%	9%	99%	5%	12%	100%
Average Bias	$T = 0$	0.137	0.155	-0.178	0.076	0.219		0.076	0.146	-0.269	0.288	0.128	-0.052	0.314	0.156	-0.044	0.126	0.145	-0.150
	$T = 1$	0.052	0.023	0.115	0.025	0.012		0.025	<0.001	0.014	0.037	0.045	0.019	0.038	0.055	0.026	0.056	0.030	0.140
Average Relative Bias	$T = 0$	-0.879	-0.456	-0.279	-0.081	-1.068		-0.081	-0.448	-0.419	-2.002	-0.663	-0.080	-2.845	-1.205	-0.069	0.975	-0.902	-0.241
	$T = 1$	-0.115	-0.055	0.300	-0.058	-0.023		-0.058	0.015	0.037	-0.086	-0.117	0.051	-0.097	-0.145	0.068	-0.142	-0.059	0.378
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	25%	100%	100%	24%	99%		24%	82%	65%	49%	100%		40%	100%		55%	100%	100%
	Second Split	88%	79%		75%	72%		75%	59%	43%	97%	64%		83%	61%		79%	69%	
	Third Split	50%	83%		58%	55%		58%	37%	8%	55%	63%		68%	60%		70%	65%	
Average Bias	$T = 0$	-0.437	0.115	0.466	-0.300	0.329		-0.300	0.092	0.293	-0.387	0.044		-0.396	0.120		-0.586	0.141	-0.045
	$T = 1$	1.482	0.265	1.524	-1.845	0.676		-1.845	0.353	0.584	-1.230	0.218		-1.247	0.125		-1.935	0.314	0.057
Average Relative Bias	$T = 0$	0.075	-0.019	-0.082	0.051	-0.055		0.051	-0.015	-0.047	0.065	-0.007		0.068	-0.020		0.100	-0.023	0.008
	$T = 1$	0.468	-0.087	-0.404	0.620	-0.237		0.620	-0.117	-0.186	0.408	-0.070		0.418	-0.040		0.637	-0.106	-0.015

For the first model with the dataset imputed by the MissForest method, the results of SIDES are not comparable since it only successfully detects subgroups from one single simulation.

Table 4.4: Performance Comparisons for Scenario Three Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	73%	90%		80%	74%		80%	80%	1%	66%	79%	100%	69%	71%	100%	68%	83%	100%
	Second Split	58%	62%		32%	33%		32%	52%	0%	41%	58%	100%	31%	64%	100%	52%	58%	0%
	Third Split	41%	39%		38%	30%		38%	47%	0%	26%	38%		34%	43%	0%	25%	35%	0%
	First Split on X_1	8%	17%		13%	0%		13%	17%	0%	5%	13%	99%	4%	5%	99%	6%	17%	100%
Average Bias	$T = 0$	-0.032	-0.028		-0.172	0.064		-0.172	0.044	-0.170	0.019	0.134	0.064	-0.030	0.075	-0.002	-0.133	0.069	-0.095
	$T = 1$	0.043	0.002		-0.005	<0.001		-0.005	-0.017	0.148	0.009	-0.025	0.006	-0.033	0.020	-0.066	-0.046	0.005	0.044
Average Relative Bias	$T = 0$	-0.479	-0.508		-1.935	-0.668		-1.935	-0.387	-0.269	-0.073	-0.060	0.105	0.246	-1.075	-0.003	-0.501	-0.732	-0.159
	$T = 1$	-0.062	-0.015		0.085	0.096		0.085	0.120	0.393	0.003	0.134	0.016	0.045	-0.078	-0.178	0.288	0.179	0.110
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	27%	61%	99%	27%	57%		27%	63%	100%	30%	65%		30%	57%		31%	66%	100%
	Second Split	39%	35%	0%	36%	32%		36%	27%	0%	45%	34%		51%	35%		39%	40%	0%
	Third Split	35%	28%	0%	32%	30%		32%	20%	0%	31%	30%		31%	24%		38%	34%	
Average Bias	$T = 0$	0.038	0.058	-0.006	0.131	0.140		0.131	0.030	-0.217	0.031	0.014		0.005	0.078		0.045	0.034	-0.007
	$T = 1$	-1.686	0.234	-0.007	-1.789	0.309		-1.789	0.100	0.575	-0.530	0.284		-0.576	0.267		-1.761	0.176	<0.001
Average Relative Bias	$T = 0$	-0.006	-0.009	0.001	-0.021	-0.022		-0.021	-0.005	0.033	-0.005	-0.003		-0.001	-0.012		-0.007	-0.006	0.001
	$T = 1$	0.629	-0.060	0.002	0.700	-0.092		0.700	-0.001	-0.251	0.212	-0.088		0.240	-0.060		0.696	-0.036	<0.001

Table 4.5: Performance Comparisons for Scenario Four Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES [9]	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	73%	90%		55%	74%		55%	64%	0%	72%	83%	100%	63%	80%	100%	78%	86%	100%
	Second Split	58%	62%		30%	36%		30%	34%	0%	46%	46%	100%	35%	46%	100%	38%	59%	0%
	Third Split	41%	39%		60%	31%		60%	38%	0%	28%	37%	0%	29%	33%	100%	45%	40%	0%
	First Split on X_1	8%	17%		0%	0%		0%	11%	0%	4%	5%	99%	4%	9%	99%	10%	26%	100%
Average Bias	$T = 0$	-0.032	-0.028		-0.024	0.064		-0.024	-0.006	-0.012	0.048	0.112	-0.006	-0.025	0.132	0.002	-0.057	-0.004	-0.097
	$T = 1$	0.043	0.002		-0.011	<0.001		-0.011	-0.006	-0.036	0.014	0.016	-0.072	-0.037	-0.002	-0.060	0.038	-0.003	0.062
Average Relative Bias	$T = 0$	-0.479	-0.508		-0.072	-0.668		-0.072	0.258	-0.020	-0.549	-1.168	-0.011	0.042	-0.894	0.003	-1.022	-0.193	-0.161
	$T = 1$	-0.062	-0.015		0.031	0.096		0.031	0.039	-0.010	-0.104	-0.151	-0.194	0.025	-0.160	-0.163	-0.039	-0.001	0.157
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	27%	61%	99%	24%	57%		24%	55%	0%	32%	67%		31%	71%		32%	71%	99%
	Second Split	39%	35%	0%	25%	32%		25%	30%	97%	48%	36%		49%	23%		34%	27%	0%
	Third Split	35%	28%	0%	29%	30%		29%	38%	0%	30%	22%		31%	35%		35%	33%	
Average Bias	$T = 0$	0.038	0.058	-0.006	0.370	0.140		0.370	-0.012	-0.247	0.005	0.013		-0.044	0.098		0.034	0.099	-0.006
	$T = 1$	-1.686	0.234	-0.007	-1.838	0.309		-1.838	-0.044	-0.502	-0.680	0.287		-0.656	0.260		-1.798	0.264	0.016
Average Relative Bias	$T = 0$	-0.006	-0.009	0.001	-0.061	-0.022		-0.061	0.001	0.041	<0.001	-0.002		0.008	-0.015		-0.005	-0.017	0.001
	$T = 1$	0.629	-0.060	0.002	0.733	-0.092		0.733	0.032	0.184	0.264	-0.063		0.253	-0.067		0.682	-0.066	-0.004

[9] For the first model with the dataset obtained by using the listwise deletion method, SIDES only detects subgroups from one simulation.

Table 4.6: Performance Comparisons for Scenario Five Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	100%	100%	100%	93%	79%		93%	83%	69%	99%	96%	99%	95%	99%	99%	98%	100%	100%
	Second Split	100%	100%	100%	95%	3%		95%	79%	36%	85%	98%	100%	83%	98%	100%	98%	100%	100%
	Third Split	75%	81%	0%	70%	16%		70%	66%	46%	79%	85%		71%	83%		92%	85%	0%
	First Split on X_1	0%	11%	100%	1%	0%		1%	40%	36%	5%	23%	99%	1%	15%	99%	0%	6%	100%
Average Bias	$T = 0$	0.137	0.155	-0.178	0.066	0.131		0.066	-0.004	-0.119	0.199	0.138	-0.074	0.224	0.103	-0.069	0.053	0.138	-0.073
	$T = 1$	0.052	0.023	0.115	0.061	0.071		0.061	0.006	0.159	0.055	-0.003	0.029	0.073	0.020	0.025	0.054	0.010	0.166
Average Relative Bias	$T = 0$	-0.879	-0.456	-0.279	-0.218	-0.199		-0.218	-0.006	-0.191	-1.565	-0.890	-0.118	-0.738	-0.018	-0.109	-0.251	-0.059	-0.115
	$T = 1$	-0.115	-0.055	0.300	-0.159	-0.193		-0.159	0.018	0.425	-0.142	0.014	0.081	-0.185	-0.048	0.070	-0.127	-0.026	0.455
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	25%	100%	100%	30%	97%		30%	95%	33%	24%	100%		19%	100%		32%	99%	
	Second Split	88%	79%		80%	12%		80%	60%	100%	90%	68%		95%	71%		77%	72%	
	Third Split	50%	83%		52%	6%		52%	57%	100%	75%	72%		62%	61%		100%	67%	
Average Bias	$T = 0$	-0.437	0.115	0.466	-0.463	0.282		-0.463	0.186	0.113	-0.311	0.048		-0.183	0.095		-0.458	0.140	
	$T = 1$	1.482	0.265	1.524	-1.870	0.557		-1.870	0.431	0.430	-0.954	0.165		-0.799	0.300		-1.619	0.264	
Average Relative Bias	$T = 0$	0.075	-0.019	-0.082	0.078	-0.047		0.078	-0.031	-0.018	0.054	-0.008		0.032	-0.016		0.078	-0.023	
	$T = 1$	0.468	-0.087	-0.404	0.635	-0.197		0.635	-0.146	-0.153	0.331	-0.056		0.275	-0.096		0.547	-0.089	

Table 4.7: Performance Comparisons for Scenario Six Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	100%	100%	100%	82%	79%		82%	89%	100%	82%	99%	99%	93%	98%	99%	100%	100%	100%
	Second Split	100%	100%	100%	60%	3%		60%	79%	0%	74%	89%	100%	74%	88%	100%	93%	99%	99%
	Third Split	75%	81%	0%	48%	16%		48%	54%	100%	53%	88%	0%	57%	85%	100%	85%	80%	0%
	First Split on X_1	0%	11%	100%	7%	0%		7%	20%	100%	16%	51%	99%	23%	31%	99%	1%	6%	100%
Average Bias	$T = 0$	0.137	0.155	-0.178	0.081	0.131		0.081	0.135	-0.088	0.095	0.058	-0.036	0.087	0.048	-0.001	-0.014	0.104	-0.094
	$T = 1$	0.052	0.023	0.115	0.006	0.071		0.006	0.012	0.188	0.027	-0.016	0.085	0.056	0.011	0.089	0.021	0.029	0.069
Average Relative Bias	$T = 0$	-0.879	-0.456	-0.279	-1.084	-0.199		-1.084	0.070	-0.138	-1.259	-0.341	-0.057	-0.327	-0.390	-0.002	-0.946	-0.267	-0.149
	$T = 1$	-0.115	-0.055	0.300	0.001	-0.193		0.001	-0.048	0.499	-0.070	0.050	0.237	-0.147	-0.032	0.251	-0.055	-0.067	0.191
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	25%	100%	100%	28%	97%		28%	67%	68%	42%	100%		61%	100%		50%	100%	
	Second Split	88%	79%		48%	12%		48%	37%	0%	100%	72%		89%	74%		82%	66%	
	Third Split	50%	83%		42%	6%		42%	42%	0%	55%	66%		67%	63%		83%	52%	
Average Bias	$T = 0$	-0.437	0.115	0.466	-0.050	0.282		-0.050	0.095	-0.110	-0.406	0.038		-0.340	0.064		-0.680	0.195	
	$T = 1$	1.482	0.265	1.524	-1.840	0.557		-1.840	0.298	-0.028	-1.199	0.283		-1.005	0.255		-1.920	0.290	
Average Relative Bias	$T = 0$	0.075	-0.019	-0.082	0.008	-0.047		0.008	-0.016	0.018	0.068	-0.006		0.058	-0.011		0.115	-0.032	
	$T = 1$	0.468	-0.087	-0.404	0.653	-0.197		0.653	-0.099	0.015	0.401	-0.096		0.339	-0.089		0.636	-0.101	

Table 4.8: Performance Comparisons for Scenario Seven Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES [16]
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	73%	90%		68%	38%		68%	71%		64%	72%	99%	58%	77%	100%	73%	82%	100%
	Second Split	58%	62%		44%	8%		44%	45%		35%	55%	100%	40%	46%	100%	39%	49%	0%
	Third Split	41%	39%		40%	27%		40%	41%		24%	43%	100%	20%	39%		52%	40%	0%
	First Split on X_1	8%	17%		5%	0%		5%	15%		41%	15%	99%	2%	15%	99%	15%	12%	100%
Average Bias	$T = 0$	-0.032	-0.028		-0.013	0.121		-0.013	0.006		0.045	0.188	0.039	0.027	0.192	0.064	-0.159	0.024	-0.095
	$T = 1$	0.043	0.002		-0.016	0.017		-0.016	-0.022		-0.005	0.010	-0.027	-0.009	0.049	0.002	-0.028	0.007	0.062
Average Relative Bias	$T = 0$	-0.479	-0.508		-1.018	-1.339		-1.018	-0.131		-0.326	-0.176	0.063	-0.055	-0.584	0.105	-0.153	-0.590	-0.159
	$T = 1$	-0.062	-0.015		-0.043	-0.026		-0.043	0.139		0.062	0.009	-0.072	0.104	-0.123	0.005	0.073	0.007	0.157
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	27%	61%	99%	27%	47%		27%	58%	66%	34%	67%		32%	62%		35%	65%	0%
	Second Split	39%	35%	0%	30%	11%		30%	40%	33%	47%	25%		47%	24%		42%	38%	0%
	Third Split	35%	28%	0%	35%	10%		35%	29%	96%	26%	33%		28%	21%		35%	39%	0%
Average Bias	$T = 0$	0.038	0.058	-0.006	0.272	0.134		0.272	-0.046	-0.094	0.011	0.122		0.001	0.183		0.009	0.046	0.035
	$T = 1$	-1.686	0.234	-0.007	-1.803	0.308		-1.803	0.078	-0.018	-0.626	0.245		-0.613	0.377		-1.791	0.291	-0.683
Average Relative Bias	$T = 0$	-0.006	-0.009	0.001	-0.045	-0.022		-0.045	0.007	0.015	-0.001	-0.020		<0.001	-0.031		-0.001	-0.007	-0.005
	$T = 1$	0.629	-0.060	0.002	0.707	-0.095		0.707	-0.027	0.025	0.247	-0.051		0.257	-0.092		0.670	-0.090	0.189

For the second model with the dataset imputed by the MissForest method, SIDES only detects subgroups from one single simulation, so the results are not comparable.

Table 4.9: Performance Comparisons for Scenario Eight Based on 100 Simulations

		Complete Dataset			Dataset with Missing			Listwise Deletion			MIPMM			MIEM			MissForest		
		QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES	QUINT	IT	SIDES [19]
Model One: Tree Structure with Continuous Outcomes																			
Variable Selection	First Split	73%	90%		44%	37%		44%	37%		39%	65%	100%	46%	59%	100%	82%	83%	
	Second Split	58%	62%		40%	8%		40%			25%	26%	100%	19%	34%	100%	42%	43%	
	Third Split	41%	39%			27%					22%	29%		24%	34%	100%	39%	46%	
	First Split on X_1	8%	17%		6%	0%		6%	6%		5%	20%	99%	8%	11%	100%	5%	25%	
Average Bias	$T = 0$	-0.032	-0.028		0.040	0.121		0.040	-0.015		-0.041	0.111	0.016	0.003	0.095	0.035	-0.026	-0.009	
	$T = 1$	0.043	0.002		-0.205	0.017		-0.205	0.033		0.024	0.017	-0.064	-0.048	0.045	-0.046	-0.034	-0.002	
Average Relative Bias	$T = 0$	-0.479	-0.508		-0.491	-1.339		-0.491	-0.313		0.233	-1.691	0.025	-0.377	-0.907	0.057	-0.406	-0.308	
	$T = 1$	-0.062	-0.015		0.517	-0.026		0.517	0.592		-0.043	-0.014	-0.174	0.090	-0.102	-0.124	0.112	0.026	
Model Two: Linear Regression with Continuous Outcomes																			
Variable Selection	First Split	27%	61%	99%	37%	47%		37%	32%		30%	58%		30%	60%		34%	63%	0%
	Second Split	39%	35%	0%	32%	11%		32%	0%		46%	23%		44%	22%		33%	36%	0%
	Third Split	35%	28%	0%	26%	10%		26%	0%		25%	17%		31%	26%		40%	37%	0%
Average Bias	$T = 0$	0.038	0.058	-0.006	0.377	0.134		0.377	-0.003		0.013	0.093		-0.006	0.089		0.068	0.057	0.035
	$T = 1$	-1.686	0.234	-0.007	-1.596	0.308		-1.596	-0.097		-0.610	0.212		-0.687	0.243		-1.716	0.208	-0.683
Average Relative Bias	$T = 0$	-0.006	-0.009	0.001	-0.060	-0.022		-0.060	<0.001		-0.002	-0.016		0.001	-0.015		-0.011	-0.010	-0.005
	$T = 1$	0.629	-0.060	0.002	0.642	-0.095		0.642	0.022		0.247	-0.062		0.286	-0.084		0.644	-0.056	0.189

For the second model with the datasets imputed by the MissForest method, SIDES method only detects subgroups from one simulation.

Chapter 5

Application Study

5.1 Study Design

We applied our three methods to a non-small cell lung cancer (NSCLC) dataset obtained from a retrospective observational study. The data contains 166 patients, among whom 153 were diagnosed with adenocarcinoma. Since the number of patients with adenocarcinoma dominates the dataset, we only analysed these 153 patients to avoid any potential bias caused by the imbalance of the histological results.

For these 153 patients, we obtained their patient characteristics at diagnosis and clinical characteristics. Patient characteristics included age, gender, and number of pack-years they smoked, while the clinical characteristics contained cancer stage, systemic therapy information, surgery, radiation, progression, recurrence, vital status, and survival month (only for the patients who died). All patients showed a positive result to the epidermal growth factor receptor (EGFR) mutations test, so we excluded this factor from the clinical characteristics. The mean age of the 153 patients was 65.42 years, with 65% of them female and 49% of them smokers. The number of pack-years smoked contains missing values, with a missing proportion equal to 12%. Based on the complete observations, the average number of pack-years equals to 12.51. Most patients were at an advanced stage of cancer (78%), and received ra-

diation (63%) and systemic therapy (86%) treatments. A small number of patients received surgery (17%). Among all these patients, 33% showed a progression and 21% suffered from a recurrence of cancer. 112 (73%) patients died during the range of the study year and the average survival months was 25 months for deceased individuals. The detailed information is shown in Table 5.1.

Table 5.1: Candidate Covariates in NSCLC Study

Covariates	**Description**	**Values/(Min., Max.)**
Patient Characteristics at Diagnosis		
Age	Patient's age at the diagnosis date	65.43 (32.31, 87.53)
Gender	Patient's gender	0 (Male, 53), 1 (Female, 100)
Packs	Number of pack-years	12.51 (0, 150)
Clinical Characteristics		
Stage	Combine 7 stages into 2 stages(IA-IIB, IIIA-IV)	1 (Early stage, 34), 2 (Late stage, 119)
Surgery	If patient got surgery at some point	0 (No, 127), 1 (Yes, 26)
Radiation	If patient got radiation therapy at some point	0 (No, 57), 1 (Yes, 96)
Systemic Therapy	Based on first line chemotherapy regimens	0 (No, 21), 1 (TKI, 111), 2 (others, 21)
Progression	Based on the first progression date	0 (No, 102), 1 (Progression, 51)
Recurrence	Based on the first recurrence date	0 (No, 121), 1 (Recurrence, 32)
Vital Status	Vital status	0 (Alive, 38), 1 (Dead, 112), NA (3)
Survival Month	Survival months for patients who died	25.11 (0.60, 124.93)

Given that many patients received several different systemic therapy regimens during the follow-up time period, we only considered the first line regimen as our treatment variable for simplification purpose. For the first line therapy, patients received one of the four different regimens: TKI (111 patients), cytotoxic (20 patients), trial (1 patient) and no systemic therapy (21 patients). Based on these categories, we evaluated two different treatment

conditions: our first treatment condition (T_1) represents the patients who got a systemic therapy (cytotoxic, TKI, or trial: $T_1 = 1$) and the patients who did not receive any systemic therapy ($T_1 = 0$). The second treatment condition (T_2) represents patients under TKI treatment ($T_2 = 1$) and patients under other systemic therapy treatments ($T_2 = 0$). The second treatment variable is used to evaluate if the patients under TKI treatment would have a better prognosis compared to the patients under other treatments. Twenty-one patients who did not receive any systemic therapy were excluded from the second treatment condition.

For the outcome variables, we mainly evaluated the relationship between the patients' survival month and their baseline and clinical characteristics. Based on the patient's vital status, we had 38 patients still alive, 112 who died and 3 with unknown status. The survival month information was only available for those 112 patients who died. The following sections show the application results based on this outcome variable. The results based on other outcome variables (progression, recurrence and vital status) are shown in Appendix B.

5.2 Parameters for the Methods

Before jumping into the results of the application study, we list here the parameters we used when applying these three methods in the software program, R.

For the IT methods, we used the R functions provided by Su et al. (2009) with the following parameters: (1) the minimum number of observations for determining terminal nodes ($min.ndsz$) is set as 20; (2) the minimum number of observations for each child node ($n0$) is equal to 10; and (3) the maximum height of a tree ($max.depth$) is specified as 10.

We adopted the following options when applying the QUINT method (Dusseldorp and Van Mechelen, 2014): (1) at the splitting stage, we used the difference in means to estimate the differential effect; (2) the minimum absolute standardized mean difference ($dmin$) is equal to 0; and (3) the maximum number of leaves ($maxl$) is specified as 10.

The following parameters choices were applied to the SIDES method (Lipkovich et al.,

2011): (1) for the splitting criterion ($num - crit$), we used the third method mentioned in Chapter 3.3.1, which is $p_3 = \max(p_1, p_2)$ (Lipkovich et al., 2011); (2) the maximum number of best promising subgroups (S) is pre-set as 5; (3) the maximum number of levels (L) for a subgroup is specified as 3; (4) the relative improvement parameter γ is equal to 1; (5) the number of permutations for the subgroup confirmation is 200; (6) $\alpha = 0.05$; and (7) the minimum desired subgroup size is 20.

5.3 Results

This section shows the results of applying IT, QUINT, and SIDES to the NSCLC data after addressing missing values. Since survival month data are only available for patients who died, we have a relatively small sample size, with 112 patients in the first treatment condition, and 98 patients in the second condition. Some methods failed to detect any subgroup because some categories ended up with only one observation, which make finding the subgroup harder. For the methods that did work, the results are shown in the following sections.

5.3.1 The First Treatment Condition

For the first treatment condition, the patients' characteristics (Table 5.2) show that patients who received treatment were younger than patients who did not receive systemic therapy. The results from IT also confirm that patients' age (75.1) is an important splitting variable to divide subgroups regarding to patients' survival month. Both IT and SIDES found that subgroups divided by the different cancer stage show an interaction effect with the systemic therapy treatment. It indicates that patients at an early stage of cancer experienced different treatment effects from systemic therapy compared to the patients at an advanced stage, especially when taking into account their age, gender or the number of pack-years they smoked. SIDES also found that the number of pack-years is an important factor which can

be used to obtain subgroups with differential treatment effects (Table 5.3).

Table 5.2: Patient Characteristics in Two Treatment Groups: Survival Month Outcome with T_1

		$T_1 = 0$ ($n = 14$)		$T_1 = 1$ ($n = 98$)	
		No.	%	No.	%
Age	Mean (SD)	72.95	12.63	64.73	12.27
Packs	Median (IQR)	4.60	(0.04, 25.00)	0.00	(0.00, 20.00)
Gender	Male	5	35.71	38	38.78
	Female	9	64.29	60	61.22
Stage	Early Stage	4	28.57	9	9.18
	Late Stage	10	71.43	89	90.82
Surgery	Yes	1	7.14	9	9.18
Surgery	No	13	92.86	89	90.82
Radiation	Yes	6	42.86	68	69.39
	No	8	57.14	30	30.61
Progression	Yes	1	7.14	35	35.71
	No	13	92.86	63	64.29
Recurrence	Yes	1	7.14	16	16.33
	No	13	92.86	82	83.67
Outcome Variables					
Survival Month	Median (IQR)	7.65	(3.10, 11.43)	18.70	(11.26, 34.19)

Table 5.3: Results for the First Treatment Condition

Treatment Effect Method	**Imputation Method**	**Subgroup**
IT	MIEM	{ Stage = 1 and Age $\leq$ 75.01 }

Table 5.3: Results for the First Treatment Condition

Treatment Effect Method	Imputation Method	Subgroup
IT	MIEM	{ Stage = 2 and Gender = 1 }
	MIPMM	{ Stage = 1 and Age $\leq$ 75.01 }
		{ Stage = 2 and Gender = 1 }
SIDES	MIEM	{ Stage = 1 }
		{ Packs $\leq$ 26 }
		{ Packs $\leq$ 20 }
		{ Stage = 1 and Packs $\leq$0 }
	MIPMM	{ Stage = 1 }
		{ Packs $\leq$ 26 }
		{ Packs $\leq$ 18 }
		{ Stage = 1 and Packs $\leq$ 0 }

5.3.2 The Second Treatment Condition

The characteristics of patients under the second treatment condition (T_2: TKI *vs.* other therapies) show that the number of patients who received TKI treatment was six times more than the number of patients who received other systemic therapies (Table 5.4). However, patients who received other systematic therapies treatment were younger than patients who received TKI treatment. For the results, in addition to the factors that we found from the first treatment condition, SIDES also found that subgroups divided by the patients' age combined with their recurrence, progression, or surgery status show differential treatment effects from the TKI therapy compared to the other groups of patients (Table 5.5).

Table 5.4: Patient Characteristics in Two Treatment Groups: Survival Month Outcome with T_2

		$T_2 = 0$ ($n = 15$)		$T_2 = 1$ ($n = 83$)	
		No.	%	No.	%
Age	Mean (SD)	59.85	9.43	65.61	12.56
Packs	Median (IQR)	0.00	(0.00, 25.00)	0.00	(0.00, 20.00)
Gender	Male	5	33.33	33	39.76
	Female	10	66.67	50	60.24
Stage	Early Stage	2	13.33	7	8.43
	Late Stage	13	86.67	76	91.57
Surgery	Yes	12	80.00	77	92.77
	No	3	20.00	6	7.23
Radiation	Yes	1	6.67	29	34.94
	No	14	93.33	54	65.06
Progression	Yes	12	80.00	51	61.45
	No	3	20.00	32	38.55
Recurrence	Yes	9	60.00	73	87.95
	No	6	40.00	10	12.05
Outcome Variables					
Survival Month	Median (IQR)	37.43	(19.22, 63.37)	17.70	(10.73, 27.72)

Table 5.5: Results for the Second Treatment Condition

Treatment Effect Method	**Imputation Method**	**Subgroup**
IT	MIEM	{ Stage = 1 and Age $\leq$ 73.60 }
	MIPMM	{ Stage = 1 and Age $\leq$ 73.60 }
QUINT	MIEM/MIPMM/MissForest	{ Age $\leq$ 51.43 }

Table 5.5: Results for the Second Treatment Condition

Treatment Effect Method	**Imputation Method**	**Subgroup**
QUINT	MIEM/MIPMM/MissForest	{ $51.43 \leq$ Age ≤ 71.72 }
SIDES	MIEM/MIPMM	{ Surgery $= 1$ }
		{ Age ≤ 49.35 }
		{ Surgery $= 1$ and Age ≤ 51.84 }
		{ Age ≤ 49.35 and Recurrence $= 0$ }
		{ Age ≤ 49.35 and Progression $= 0$ }
		{ Age ≤ 49.35 and Stage $= 2$ }

5.4 Conclusions

All things considered, the most frequently selected covariates were patient age and the number of pack-years they smoked. For the patient age, most of the results show that it is an important splitting variable to be used to divide subgroups, independently or combined with other covariates. Considering that patients who received treatment were younger than patients who did not receive any systemic therapy, we believe that what this factor actually reflects is the selection for being delivered a treatment. Since the study did not randomize the patients into two groups, age might be a factor that has impact on which treatments these patients received, rather than the performance of the treatment itself.

Our results also show that the subgroups partitioned by the number of pack-years of smoking show a differential treatment effect, especially for the first treatment condition (systemic therapy *vs.* no systemic therapy). Since smoking is a well-known risk factor for the NSCLC, it is not a surprise that the results show a connection between the smoking status and the patients' outcome (Molina et al., 2008). However, the patient characteristics show that under the first treatment condition, there are more smokers in the control group

compared to the systemic therapy group. From this point of the view, smoking status might be a reflection of the patients' treatment condition. It can also explain why the number of pack-years is selected less frequently under the second treatment condition, which has a relatively balanced number of smokers in the two arms.

In addition to the previous factors, subgroups divided by the patients' age combined with their cancer stage, surgery, recurrence or progression status also show an interaction effect with the systemic therapy and TKI treatment. These identified subgroups might be useful for further exploring. However, are they real subgroups? We do not know for sure. Due to the unbalanced patient characteristics and sample sizes in the two groups, it is hard to draw any solid conclusion.

In terms of the model performances, similar to what we found from the simulation study, the results for IT and QUINT are consistent in some ways. However, the results of SIDES are slightly different from that of the other two methods. Furthermore, the performance of the QUINT method is less stable compared to the IT method, since this method failed to detect subgroups in the first treatment condition. It indicates that the QUINT tends to fail when there is an increased imbalance in the number of patients between the two groups. The SIDES method tends to list all the potential subgroups with different combinations of the covariates, which makes the results less convincing. This finding also supports the results from the simulation study, in that the QUINT and the SIDES methods are less preferable compared to the IT method.

For the methods addressing missing data, MIPMM and MIEM perform equally well and their results are comparable. The performance of the MissForest is less satisfying for this dataset. This result echoes what we got from the simulation study in that the MissForest method tends to perform better when the dataset is generated from a tree-based structure. An assumption of the data structure needs to be made during the method selection, although in our data, the tree structure is unknown.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this study, we focused on the treatment effect models to assess predictive biomarkers or other factors that provide information about the effect of a therapeutic intervention. We evaluated and addressed the consequences of having missing data when using three treatment effect models: IT, QUINT, and SIDES. Simulation studies were used to compare the model performance for treatment effects when using complete, incomplete and imputed data across various scenarios when the sample size, proportion of missingness and number of missing variables differ. We also applied these three methods to a NSCLC dataset obtained from a retrospective study.

In sum, both the simulation and application studies indicated that the IT and QUINT methods are equivalent in most situations. However, in situations such as the existence of a severe imbalance between two groups, it would be better to use the IT method. The SIDES method requires a descent sample size to achieve a better partitioning outcome. Furthermore, the results are less comparable to the other two methods. This difference is due to the varied mechanisms of these methods. SIDES is a local treatment effect model and tends to search for one subgroup with a maximum treatment effect, while IT and QUINT are the global

treatment effect models and aim to divide the population into several exclusive subgroups. The diverse purposes of the three methods lead to the different results between the models, so the study objective need to be considered when choosing these methods.

For the missing data, both studies show that the assumption of the data structure plays an important part in the choice of the data completion methods. In most cases, it would be a safe option to choose multiple imputation methods, such as MIPMM and MIEM. For some rare situations when the data have a tree-based structure, MissForest would be a perfect choice. For a tree-based structure dataset, addressing the missing data by MissForest will boost the performance of the treatment effect models. It is because the treatment effect models and the MissForest are sharing a similar underlying methodology. The listwise deletion would be the last choice for handling missing data.

One difficulty in the real world application is to evaluate the data structure. There are two potential ways to do it. The first one is making decision based on the clinical expert opinion. For example, we might be able to consider some genetic data as following a tree-based structure if there is a clinical evidence to show that. The second way is applying sensitivity analysis to compare the results from the tree-based methods and the regression-based methods. Based on the results, it is relatively easy to evaluate which method is a good fit.

6.2 Future work

For future work, these methods could be applied to a real randomized dataset. Due to the limitations of the available dataset, it is hard to draw any solid conclusion for the application study. A result obtained from a balanced dataset would be more convincing. Besides, equal numbers of patients in each group and a larger sample size might also improve the performance of the treatment effect methods, especially for the QUINT and SIDES methods.

In the situation that only observational data are available, we could adopt some methods

to balance the baseline characteristics in each group to eliminate the potential treatment assignment difference. Such methods include propensity score (PS) matching, inverse probability of treatment weighting (IPTW) and genetic matching (GM) (Kreif et al., 2012). However, these methods should only be applied under the assumption that no unobserved confounding exists.

Since our study mainly focuses on tree-based treatment effect methods, we did not include other types of subgroup analysis models, such as outcome modeling methods and regression-based models. The comparison of the performances among these models might shed some light on model selection in subgroup identification studies.

Given that missing data is an unavoidable problem in the real-life data application, it would be a great idea to introduce some other novel methods to address missing data to overcome limitation of the existing methods. We could also consider more restrictive assumption regarding to the missingness, such as MNAR.

Bibliography

Alemayehu, D., Chen, Y., and Markatou, M. (2017). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research*, pages 1–21.

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., and Yue, L. (2015). Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research*, 7, 286–303.

Alosh, M., Huque, M. F., Bretz, F., and D'Agostino, R. B. (2017). Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in Medicine*, 36, 1334–1360.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees*. New York, NY: Chapman and Hall/CRC.

Brookes, S. T., Whitley, E., Peters, T. J., Mulheran, P. A., Egger, M., and Davey Smith, G. (2001). Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technology Assessment*, 5, 1–56.

Collins, R. (1996). Large-scale randomized evidence: trials and overviews. *Oxford Textbook of Medicine*, 48, 21–32.

Dong, Y. and Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *Springer Plus*, 2, 222.

Dusseldorp, E., Doove, L., and Van Mechelen, I. (2016). Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. *Behavior Research Methods*, 48, 650–663.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, 33, 219–237.

Ghorbani, S. and Desmarais, M. C. (2017). Performance comparison of recent imputation methods for classification tasks over binary data. *Applied Artificial Intelligence*, 31, 1–22.

Goldberger, J. J. and Buxton, A. E. (2013). Personalized medicine vs guideline-based medicine. *JAMA*, 309, 2559–2560.

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G., Steyerberg, E. W., et al. (2013). Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ*, 346, 1–11.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54, 561–581.

Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45, 1–47.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions. *Journal Clinical Oncology*, 29, 4718.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.

Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., and Sekhon, J. S. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making*, 32, 750–763.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88, 457–467.

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10, 422–439.

Lipkovich, I. and Dmitrienko, A. (2014a). *Biomarker identification in clinical trials*. New York, NY: Chapman and Hall/CRC.

Lipkovich, I. and Dmitrienko, A. (2014b). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics*, 24, 130–153.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30, 2601–2621.

Lipkovich, I., Dmitrienko, A., et al. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36, 136–196.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 14–23.

Loh, W.-Y., Fu, H., Man, M., Champion, V., and Yu, M. (2016). Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine*, 35, 4837–4855.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34, 1818–1833.

Mayer, C., Lipkovich, I., and Dmitrienko, A. (2015). Survey results on industry practices and challenges in subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research*, 7, 272–282.

Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. In *Mayo Clinic Proceedings*, volume 83, pages 584–594. Elsevier.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F. (2005). Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15, 231–239.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., and Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of Biopharmaceutical Statistics*, 26, 99–119.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21, 2917–2930.

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., et al. (2013). Prognosis research strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine*, 10, 1–9.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365, 176–186.

Ruberg, S. J. and Shen, L. (2015). Personalized medicine: Four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research*, 7, 214–229.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22, 425–446.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28, 112–118.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10, 141–158.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4, 1–26.

Sun, X., Briel, M., Walter, S. D., and Guyatt, G. H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*, 340, 850–854.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357, 2189–2194.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., and Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, 71, 645–653.

Yusuf, S., Collins, R., and Peto, R. (1984). Why do we need some large, simple randomized trials? *Statistics in Medicine*, 3, 409–420.

Zhang, H. and Singer, B. H. (2010). *Recursive partitioning and applications*. Berlin: Springer Science & Business Media.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107, 1106–1118.

Appendix A

Appendix for the Simulation Study

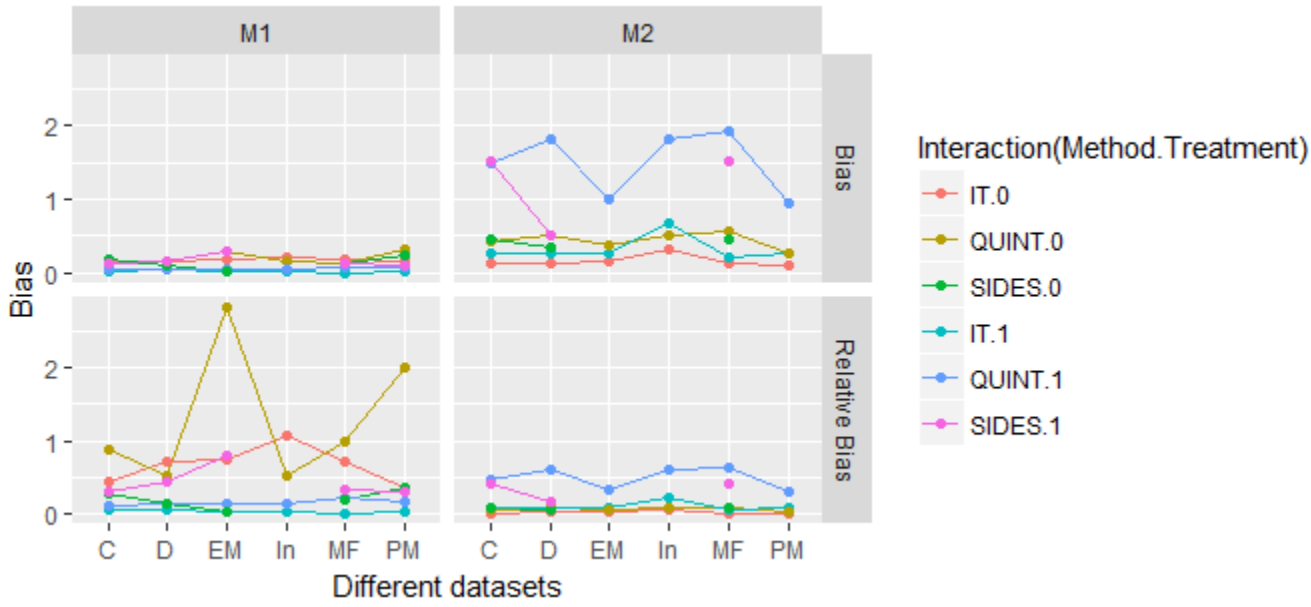

Figure A.1: The Bias and Relative Bias for Scenario One [1]

Figure A.1 is the complete plot for scenario one in the simulation study. This plot includes the results from the SIDES method for the first model when using the MissForest to deal with the missing data.

[1] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

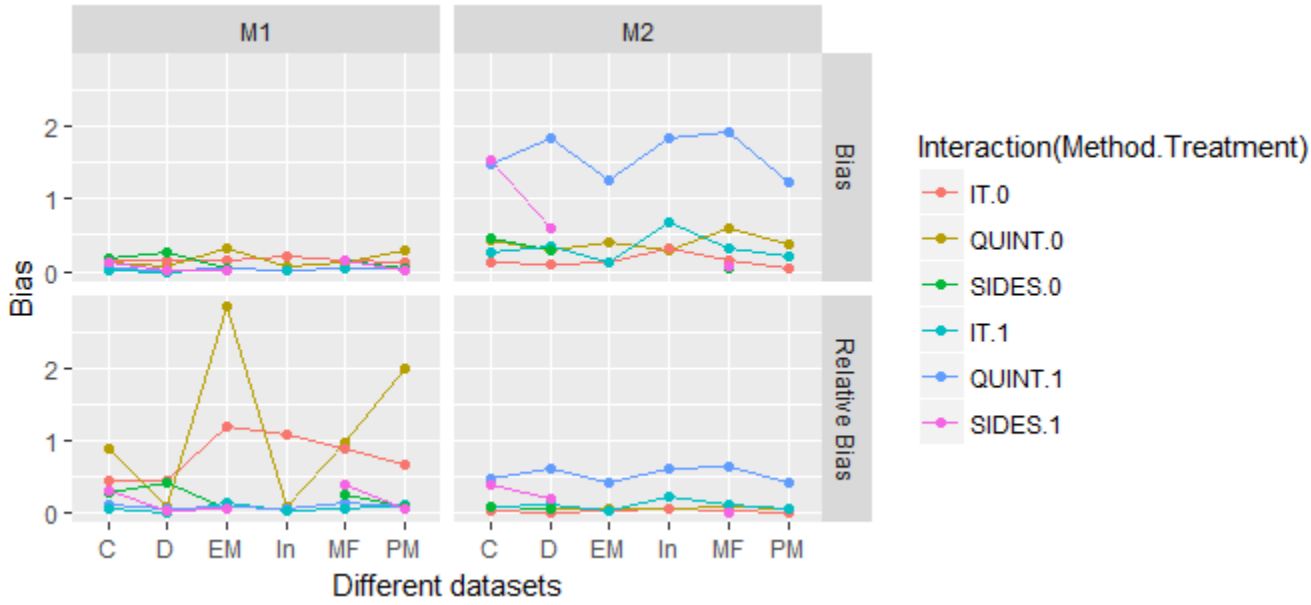

Figure A.2: The Bias and Relative Bias for Scenario Two [2]

Figure A.2 is the complete plot for scenario two in the simulation study. This plot includes the results from the SIDES method for the first model after using the MissForest to deal with the missing data.

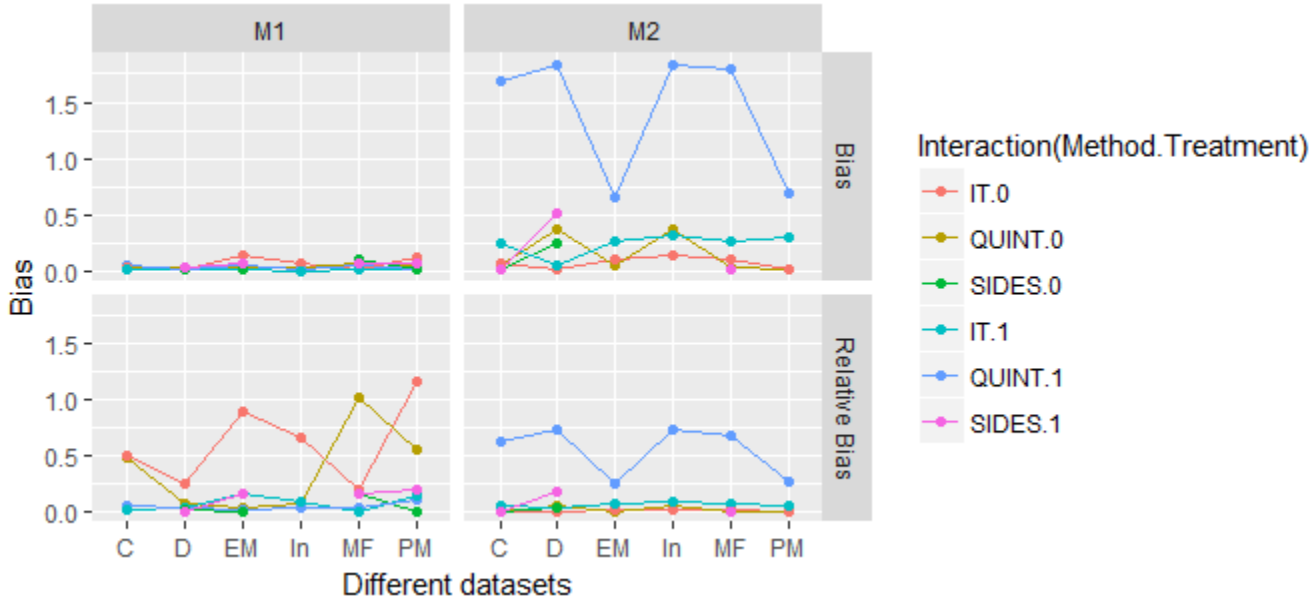

Figure A.3: The Bias and Relative Bias for Scenario Five [3]

[2]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

Figure A.3 is the complete plot (including the SIDES results for the first model after using the listwise deletion to deal with the missing data) for scenario four in the simulation study.

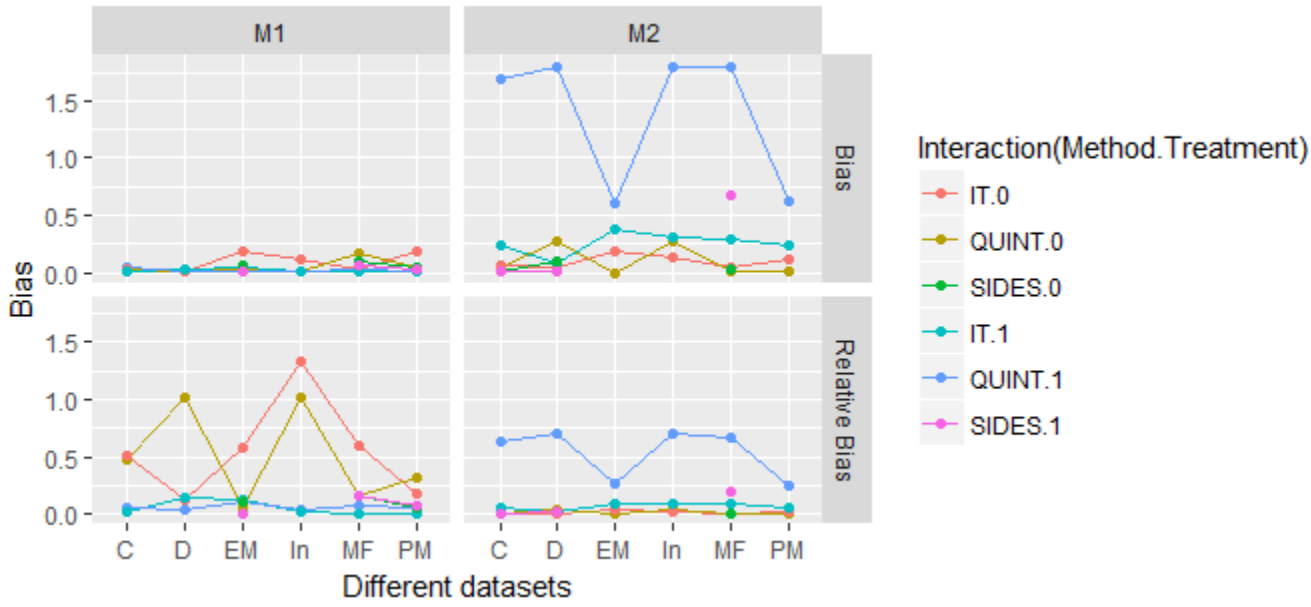

Figure A.4: The Bias and Relative Bias for Scenario Six [4]

Figure A.4 is the complete plot (including the SIDES results for the second model after using the MissForest to deal with the missing data) for scenario seven in the simulation study.

[3]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

[4]C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

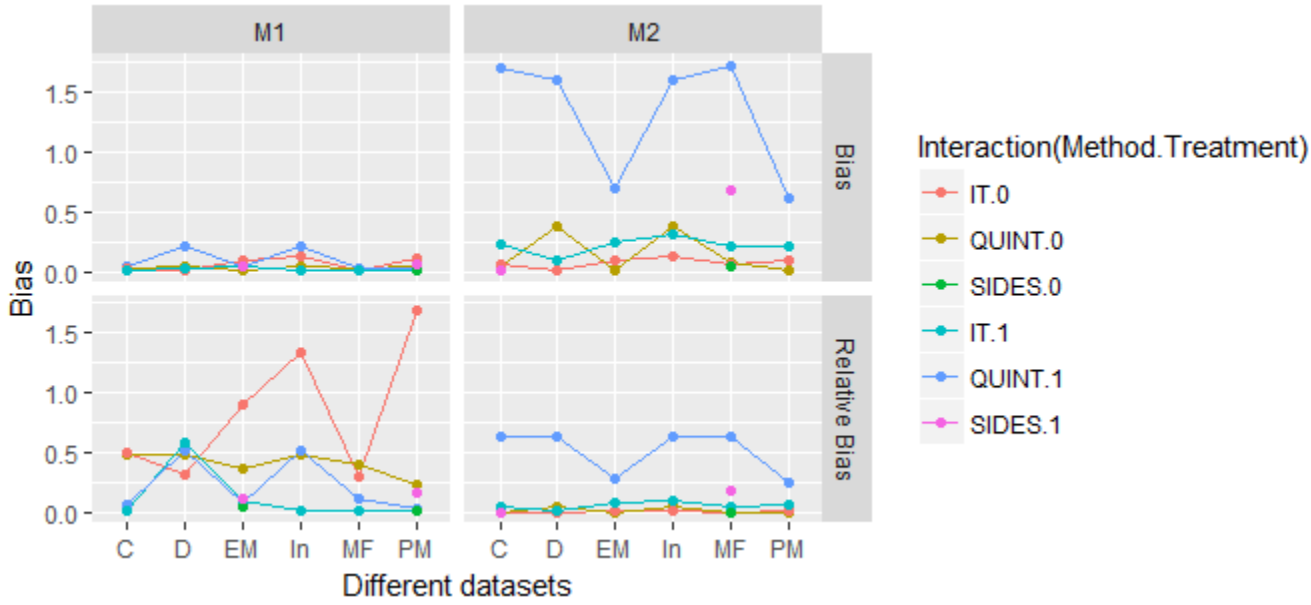

Figure A.5: The Bias and Relative Bias for Scenario Six [5]

Figure A.5 is the complete plot (including the SIDES results for the second model after using the MissForest to deal with the missing data) for scenario eight in the simulation study.

[5] C: The complete dataset without any missing data; D: Listwise deletion; EM: MIEM; In: The incomplete dataset containing missing values; PMM: MIPMM

Appendix B

Appendix for the Application Study

In addition to the survival month outcome, we also considered three outcome variables from two different stages in the NSCLC study. The first stage is in the middle of the study, as we would like to evaluate if there is any characteristic that influences patients' progression and recurrence status. The second stage is at the end of the study, where we evaluated the relationship between the patients' vital status and their baseline and clinical characteristics. For the patient's vital status, we had 38 patients still alive, 112 who died and 3 with unknown status. The three patients whose vital statuses were unknown were excluded from the analysis. The following sections show the application results for the SIDES method based on different outcome variables.

B.1 Results for the Mid-stage Outcomes

This section presents the results of applying SIDES to the mid-stage data after addressing missing values. At this point, we had two outcome variables to evaluate patients' response to a treatment, which are progression status and recurrence status.

B.1.1 Progression

In terms of the patients' progression of the disease, we evaluated their performances based on two different treatment conditions. For the first treatment condition (T_1: systemic therapy *vs.* no systemic therapy), SIDES fails to detect any subgroup. The characteristics of patients under the second treatment condition (T_2: TKI *vs.* other therapies) show that the number of patients who received TKI treatment was five times more than the number of patients who received other systemic therapies (Appendix Table B.1). There were more female patients and most patients were at the advanced stage of the cancer. The SIDES results for the second treatment condition shown that patients above and below certain age (∼73) would respond differently to the TKI therapy. The number of pack-years they smoked is also an important factor which can be used to obtain subgroups with differential treatment effects (Appendix Table B.2).

Table B.1: Patient Characteristics in Two Treatment Groups: Mid-stage Data with T_2

		$T_2 = 0$ ($n = 21$)		$T_2 = 1$ ($n = 111$)	
		No.	%	No.	%
Age	Mean (SD)	61.17	11.11	64.93	12.35
Packs	Median (IQR)	0.00	(0.00, 16.00)	0.00	(0.00, 20.00)
Gender	Male	8	38.10	39	35.14
	Female	13	61.90	72	64.86
Stage	Early Stage	5	23.81	19	17.12
	Late Stage	16	76.19	92	82.88
Surgery	Yes	5	23.81	17	15.32
	No	16	76.19	94	84.68
Radiation	Yes	18	85.71	68	61.26
	No	3	14.29	43	38.74
Outcome Variables					

Table B.1: Patient Characteristics in Two Treatment Groups: Mid-stage Data with T_2

		$T_2 = 0$ ($n = 21$)		$T_2 = 1$ ($n = 111$)	
		No.	%	No.	%
Progression	Yes	3	14.29	64	57.66
	No	18	85.71	47	42.34
Recurrence	Yes	8	38.10	21	18.92
	No	13	61.90	90	81.08

Table B.2: Results for the SIDES: Progression Outcome under T_2

Imputation Method	**Subgroup**
MIEM	$\{\text{Age} \leq 75.4\}$
	$\{\text{Age} \leq 73.5\}$
	$\{\text{Age} \leq 69.7 \text{ and Packs} \leq 26\}$
	$\{\text{Age} \leq 75.0 \text{ and Packs} \leq 38\}$
MIPMM	$\{\text{Age} \leq 75.4\}$
	$\{\text{Age} \leq 73.5\}$
	$\{\text{Age} \leq 75.4 \text{ and Packs} \leq 12\}$
	$\{\text{Age} \leq 69.7 \text{ and Packs} \leq 26\}$

B.1.2 Recurrence

The patients' characteristics for the recurrence outcome data are essentially the same as the progression outcome data (Appendix Table B.1 and Table B.3). For the patients' recurrence status, we also considered two different treatment conditions. For the first treatment con-

dition (systemic therapy *vs.* no systemic therapy), the patients' characteristics (Table B.3) show that patients who received treatment were younger than patients who did not receive systemic therapy. The results from the first condition confirm that patients' age ($\sim$70) is an important splitting variable to divide subgroups regarding to patients' recurrence status. In addition to the factors we found from the progression data, subgroups divided by the surgery status also show an interaction effect with the systemic therapy treatment. Patients who received surgery experienced different treatment effects from systemic therapy compared to the patients who did not receive surgery (Appendix Table B.4).

For the second condition (TKI *vs.* other therapies), subgroups divided by patients' cancer stage show an interaction effect with the TKI therapy. This interaction effect indicates that patients at different stages would have different responses to the TKI treatment in term of cancer recurrence (Appendix Table B.4).

Table B.3: Patient Characteristics in Two Treatment Groups: Progression Outcome with T_1

		$T_1 = 0$ ($n = 21$)		$T_1 = 1$ ($n = 132$)	
		No.	%	No.	%
Age	Mean (SD)	72.50	12.99	64.31	12.20
Packs	Median (IQR)	1.20	(0.00, 30.00)	0.00	(0.00, 20.00)
Gender	Male	6	28.57	47	35.61
	Female	15	71.43	85	63.39
Stage	Early Stage	10	47.62	24	18.18
	Late Stage	11	52.38	108	81.82
Surgery	Yes	4	19.05	22	16.67
	No	17	80.95	110	83.33
Radiation	Yes	10	47.62	86	65.15
	No	11	52.38	46	34.85
Outcome Variables					

Table B.3: Patient Characteristics in Two Treatment Groups: Progression Outcome with T_1

		$T_1 = 0$ ($n = 21$)		$T_1 = 1$ ($n = 132$)	
		No.	%	No.	%
Progression	Yes	1	4.76	50	37.88
	No	20	95.24	82	62.12
Recurrence	Yes	3	14.29	29	21.97
	No	18	85.71	103	78.03

Table B.4: Results for the SIDES: Recurrence Outcome

Treatment Condition	**Imputation Method**	**Subgroup**
T_1	MIEM	{ Surgery = 1 }
		{ Age $\leq$ 70.1 and Surgery = 1 }
		{ Age $\leq$ 70.1 and Surgery = 1 and Packs $\leq$ 25}
	MIPMM	{ Surgery = 1 }
		{ Stage = 1 }
		{ Age $\leq$ 70.1 and Surgery = 1 }
	MissForest	{ Surgery = 1 }
		{ Age $\leq$ 70.1 and Surgery = 1 }
T_2	MIEM	{ Stage = 1 }
		{ Age $\leq$ 77.1 and Stage = 1 }
		{ Stage = 1 and Packs $\leq$ 0 }
	MIPMM	{ Stage = 1 }
		{ Surgery = 1 }
		{ Age $\leq$ 77.1 and Stage = 1 }
	MissForest	{ Surgery = 1 and Packs $\leq$ 33 and Age $\leq$ 77.1 }

B.2 Results for the End-stage Outcome

Similar to the previous section for mid-study outcomes, this section shows the results of applying SIDES to the end-stage data after addressing missing values. At the end of the study, patients' vital status was used as the response variable, leaving all other variables (including progression and recurrence statuses) as the predictor variables.

B.2.1 Vital Status

Similar to the mid-stage results, the patients' characteristics based on vital status (Appendix Table B.5) also show that patients who received treatment were younger than patients who did not receive any systemic therapy. Our findings, that factors such as age and the number of packs smoked are important elements for dividing subgroups with differential treatment effects, are consistent with what we found from the mid-stage data. Except for those factors, subgroups partitioned by the patients' radiation status also show an interaction effect with the systemic therapy treatment (systemic therapy *vs.* no systemic therapy) (Appendix Table B.6). It indicates that in terms of the vital status, patients who received radiation had a different treatment effect from systemic therapy compared to the patients who did not receive any radiation, especially when considering their age, gender and progression status. The results for the second treatment condition (TKI *vs.* other therapies) echo what we got from the previous sections (Appendix Table B.6).

Table B.5: Patient Characteristics in Two Treatment Groups: Vital Status Outcome with T_1

		$T_1 = 0$ ($n = 21$)		$T_1 = 1$ ($n = 129$)	
		No.	%	No.	%
Age	Mean (SD)	72.50	12.99	64.34	12.24
Packs	Median (IQR)	1.20	(0.00, 30.00)	0.00	(0.00, 20.00)
Gender	Male	6	28.57	45	34.88
	Female	15	71.43	84	65.12
Stage	Early Stage	10	47.62	23	17.83
	Late Stage	11	52.38	106	82.17
Surgery	Yes	4	19.05	21	16.28
	No	17	80.95	108	83.72
Radiation	Yes	10	47.62	84	65.12
	No	11	52.38	45	34.88
Progression	Yes	1	4.76	50	38.76
Progression	No	20	95.24	79	61.24
Recurrence	Yes	3	14.29	28	21.71
	No	18	85.71	101	78.29
Outcome Variables					
Vital Status	Alive	7	33.33	31	24.03
	Dead	14	66.67	98	75.97

Table B.6: Results for the SIDES: Vital Status Outcome

Treatment Condition	**Imputation Method**	**Subgroup**
T_1	MIEM	{ Progression = 0 and Radiation = 1 }
		{ Progression = 0 and Radiation = 1 and Age $\geq$ 53.7 }

Table B.6: Results for the SIDES: Vital Status Outcome

Treatment Condition	**Imputation Method**	**Subgroup**
T_1	MIEM	{ Progression = 0 and Radiation = 1 and Gender = 1 }
		{ Progression = 0 and Gender = 1 and Packs $\leq$ 36.2 }
	MIPMM	{ Progression = 0 and Radiation = 1 }
		{ Progression = 0 and Radiation = 1 and Age $\geq$ 53.7 }
		{ Progression = 0 and Radiation = 1 and Gender = 1 }
		{ Progression = 0 and Gender = 1 and Age $\leq$ 72.7 }
T_2	MIEM	{ Age $>$ 75.5 }
		{ Age $>$ 75.5 and Recurrence = 0 }
		{ Age $>$ 75.5 and Progression = 0 }
		{ Age $>$ 75.5 and Packs $\leq$ 0 }
	MIPMM	{ Age $>$ 75.5 }
		{ Age $>$ 75.5 and Recurrence = 0 }
		{ Age $>$ 75.5 and Progression = 0 }
		{ Age $>$ 75.5 and Packs $\leq$ 5 }