Multivariate CACE Analysis with an Application to Arthritis Health Journal Study

by

Yue Ma

B.Sc., Nankai University, 2016

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2018

Approval

Name: **Yue Ma**

Degree: **Master of Science (Statistics)**

Title: **Multivariate CACE Analysis with an Application to Arthritis Health Journal Study**

Examining Committee:

Chair: Jinko Graham
Professor

Hui Xie
Senior Supervisor
Professor
Faculty of Health Sciences
&
Associate Member
Statistics and Actuarial Science

Joan Hu
Supervisor
Professor

Bohdan Nosyk
Supervisor
Associate Professor
Faculty of Health Sciences

Lawrence McCandless
Internal Examiner
Associate Professor
Faculty of Health Sciences
&
Associate Member
Statistics and Actuarial Science

Date Defended: **May 7, 2018**

Abstract

Treatment noncompliance is a common issue in randomized controlled trials that may plague the randomization settings and bias the treatment effect estimation. The complier-average causal effect (CACE) model has become popular in estimating the method effectiveness under noncompliance. Performing multiple univariate CACE analysis separately fails to capture the potential correlations among multivariate outcomes, which will lead to biased estimates and significant loss of power in detecting actual treatment effect. Motivated by the Arthritis Health Journal Study, we propose a multivariate CACE model to better account for the correlations among outcomes. In our simulation study, the global likelihood ratio test is conducted to evaluate the treatment effect which fails to control the type I error for moderate sample sizes. So, we further perform a parametric bootstrap test to address this issue. Our simulation results suggest that the Multivariate CACE model outperforms multiple Univariate CACE models in the precision of estimation and statistical power in the case of correlated multivariate outcomes.

Keywords: Multivariate CACE; Univariate CACE; non-compliance; MLE; statistical power; parametric bootstrap test

Dedication

To my beloved parents.

Acknowledgements

I would first like to thank my supervisor, Dr. Hui Xie, for his continual support, patience and guidance throughout my time at Simon Fraser.

Thank you to all members in this department, from professors to staff. I also want to extend special thanks to Dr. Joan Hu, Dr. Bohdan Nosyk and Dr. Lawrence McCandless for serving on my committee.

I want to express my deepest appreciation to my parents. Thank you for your endless love, unconditional support and understanding. You trusted me enough to make my own decisions and have been there for anything I needed.

Lastly, I would like to thank all my friends and fellow students. Thank you Mozhu Mu and Polly Wu for accompanying me and sharing happiness with me. Thank you Perry Sang and Shijia Wang for your help on this project. Thank you Shufei Ge, Grace Hsu, Zetong Li, Yan Lin, Dongmeng Liu, Michelle Thiessen, Ran Wang, Lingling Zhang, Charlie Zhou and Zhiyang Zhou for showing up in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Randomized controlled trial (RCT) is a type of study in which subjects are randomly assigned to either a treatment arm receiving some clinical intervention or a control (placebo) arm. RCT study is considered the gold standard to test the efficacy of a new treatment by comparing the outcomes after participants receive different interventions. Randomization minimizes the selection bias and eliminates the confounding and unobserved factors so that the difference in measurements we observe from different groups is only attributable to the different treatments the participants have received. Inferences based on randomization require that all participants adhere to their initial treatment assignments. However, noncompliance often occurs when the randomized experiments involve human subjects. In practice, participants may refuse to take the treatment due to side effects, inconvenience, etc. Therefore, noncompliance turns out to be an important issue as it may lead to biased estimates of actual treatment effect. The bias comes from the dilution of the treatment efficacy caused by noncompliance behavior.

Intention-to-treat (ITT) analysis and as-treated(AT) analysis are considered as two traditional approaches in evaluating the treatment effect. The ITT analysis (also known as as-randomized analysis) compares the outcomes between participants assigned to the treatment group and participants assigned to the control group, regardless of the actual receipt of treatment. Usually, the ITT analysis is the default approach for estimating treatment effect under noncompliance, which provides the unbiased estimate for use/program effectiveness. While the use effectiveness can be of critical importance in RCTs, the method effectiveness is our primary interest. The method effectiveness is the biological effect of the new treatment/drug, which is unaffected by noncompliance behaviors. The compliance rate would increase if the method effectiveness was known to be beneficial. The use effectiveness may vary for different populations, but the method effectiveness for one treatment should remain similar over different populations. Given the perfect compliance, the method effectiveness and use effectiveness can be treated equally for the entire population.

The ITT analysis tends to offer conservative estimate of the actual treatment effect under noncompliance, thus is biased for method effectiveness.

As-treated (AT) analysis, which aims at estimating the method effectiveness, compares the outcomes based on the actual receipt of treatment and ignores the initial assignment of treatment. A concern of the AT analysis under noncompliance is the violation of randomization assumption so that some unobserved factors may potentially corrupt the causal interpretation of treatment effects. Important as it has seemed to be, the noncompliance behavior is not random as it may depend on the health status and other characteristics of participants. For instance, patients with mild symptoms are more likely to drop-out of the treatment. Thus, the AT analysis also fails to provide an unbiased estimate of method effectiveness.

Under such circumstance, an alternative candidate called Complier-Average Causal Effect (CACE) analysis was introduced (LEWIS b. Sheiner and Rubin [1995]) as a remedy for estimating the method effectiveness. There are four possible types of compliance behavior discussed in CACE analysis: "always-taker", "never-taker", "complier" and "defier", which will be defined in Chapter 3 in detail.

One of the main advantages of the ITT analysis and the AT analysis over CACE could be the simplicity of calculations, especially in former times when people had to do calculations by hand. With the rapid development in computing technology, time-consuming calculation is no longer a problem for CACE analysis. Thus, CACE becomes a popular model for causal inferences.

1.2 Motivation

In efforts to better evaluate the treatment efficacy in RCT studies, many researchers take multiple measurements, thus producing multivariate outcomes. For example, if researchers want to evaluate the effect of a new drug D on breast cancer, many measurements including complete blood count (CBC) and some breast cancer tumor markers (e.g. CA 15-3) are taken as the outcomes. Previous literatures only considered the Univariate CACE model so that in the case of multivariate outcomes, multiple Univariate CACE models were applied to evaluate the treatment effect. However, the Multivariate CACE model will outperform multiple Univariate CACE models in three ways.

First of all, an important focus of CACE model is the compliance mechanism. In general, the compliance rate is not fully observable in RCTs, thus needs to be estimated from the observed outcomes and baseline covariates. If we perform multiple Univariate CACE analysis on k dimensional outcomes separately, we will get k different estimates of compliance rate. For a given population, however, the compliance behavior depends on baseline variables rather than the type of health outcomes; that is, we should have one compliance

rate for one population. Hence, having different estimates of compliance rate makes no sense and is hard to interpret in a scientific way.

Secondly, multiple Univariate CACE models fail to capture the potential correlations among multivariate outcomes. In many situations, we are not sure about the underlying correlations among outcomes, thus performing multiple univariate analysis may risk losing information about the given data. Even in the case of uncorrelated outcomes, the Multivariate CACE model still has a slight advantage over the Univariate CACE in precise estimation. In practice, the multivariate measurements are intuitively correlated based on their scientific meaning. Therefore, the Multivariate CACE is considered to be a properer candidate for better modeling multivariate health outcomes.

Thirdly, the significance of treatment effect is of great interest to researchers in many, if not all, RCT studies. Multiple univariate tests inflate both the experiment-wise type I error rate and the experiment-wise type II error rate (often called probability pyramiding [Haase and Ellis, 1987]) when there are more than one dependent measurements. Experiment-wise type I error rate (α_{EW}) is defined as the overall type I error when conducting a series of tests on dependent outcomes. And per-comparison type I error rate (α_{PC}) refers to the risk of a "false positive" occurring in an individual test on one of the dependent outcomes. Suppose there are k dependent outcomes, then $\alpha_{EW} = 1 - (1 - \alpha_{PC})^k$. As dimension k increases, the experiment-wise error rate escalates rapidly. For example, if we set α_{PC} to 0.05 and the dimension $k = 6$, α_{EW} will increase to 0.226. Similarly, the experiment-wise type II error rate, β_{EW}, runs into to the same problem. β_{EW} escalates exponentially as the number of dependent outcomes increases for fixed β_{PC} (per-comparison type II error). Multivariate analysis controls the escalation of β_{EW}, which ensures a higher experiment-wise power of the study.

Hence, multivariate analysis is properer for causal inferences when the health outcomes are multivariate. Multivariate analysis captures the potential correlations among multivariate outcomes, generates interpretable results and boosts our confidence of test results.

1.3 Literature Review

Dating back to the late 20th century, when randomized clinical trials became very popular for studies involving human subjects, researchers started to notice the issue of noncompliance. The occurrence of noncompliance could violate the assumption of the standard theories of randomization popularized by Fisher [1925]. Some researchers, like Lee et al. [1991], pretended the compliance was perfect and compared the outcomes between participants grouped by the arms to which they were randomized.

Economists were also interested in estimating causal effects and the dominant approach was based on structural equation models via instrumental variables (IV). Angrist

et al. [1996] proposed a framework for causal effects in the case of noncompliance. They made use of IV to get IV estimands and showed that IV estimands could be fit into the Rubin Causal Model (Holland [1986]). They also showed that the IV estimands is the average causal effect for compliers under some interpretable assumptions. This approach made it easier to interpret the critical assumptions needed for causal inferences in a scientific way and allowed for sensitivity analysis on violation of key assumptions in more straightly.

Imbens and Rubin [1997] developed a framework for estimating causal effects via Bayesian inferential methods to address the issue of imperfect compliance. They obtained the posterior estimation from EM algorithm and data augmentation algorithm. They also implemented their proposed approach to both discrete and continuous outcomes and compared the results of the posterior estimation, the maximum likelihood estimate (MLE) and the IV estimands. The results suggested that the MLE had comparable performance with the posterior estimation in terms of bias and root mean squared error, and both the posterior estimation and the MLE clearly outperformed standard IV estimands. The two stage IV estimator can be very robust but it fails to take into full consideration that the observed outcomes are mixtures of outcomes with different compliance types. In many clinical trials, when the sample size is moderate, the likelihood-based approach is considered more efficient to provide more powerful results.

Hirano et al. [2000] reanalyzed the study conducted by McDonald et al. [1992] to explore the efficacy of the influenza vaccine. Similar to Imbens and Rubin [1997], they obtained the causal effect via Bayesian approach. But one of the new features of this extended framework was that they took consideration of baseline covariates, which might influence the probability of receiving the treatment. In addition, they also relaxed the exclusion restriction assumption, one of the critical assumptions required by Angrist et al. [1996]. This assumption rules out the direct effect of treatment assignment on the final outcomes, which might violate the scientific meaning in some cases. This proposed framework could still work well when the exclusion restriction assumption is violated.

Connell [2009] employed the CACE approach to evaluate the intervention effect of adaptive prevention programs on the development of substance use behaviors. They studied the long-term outcomes of reducing tobacco-use from early adolescence through early adulthood. They only focused on the maximum likelihood approach based on a mixture modeling framework to identify the compliers in the control group and compare with the observed compliers in the treatment group. They also validated the meeting of assumptions required by the CACE model through their dataset and discussed the potential results due to the violation of exclusion restriction.

Stanger et al. [2011] conducted a similar research on the substance abuse issue. Previous studies suggested that children of substance abusers were more likely to suffer from behavioral/emotional problems. A new program called contingency management (CM) was designed to enhance the compliance with parent training by providing incentives. Be-

sides the basic CACE model, they also extended the CACE model by including covariates to predict the treatment outcomes and the compliance behavior. They modeled multivariate outcomes separately and found that only some outcomes were significantly different between different treatment groups. Compared to the ITT analysis, the CACE model provided stronger evidence of the treatment effect. Unfortunately, their study was limited by the small sample size.

Knox et al. [2014] studied the treatment compliance and the efficacy of a cognitive behavioral intervention for low back pain via CACE approach. CACE estimates showed greater difference in change scores from baseline compared to the conservative ITT approach, but both approaches reached the same conclusion that the treatment effect was statistically significant. They also investigated the sensitivity to the missing data via multiple imputation. Their research suffered much from the noncompliance issue: nearly half of the participants in the treatment group failed to adhere to their original assignment. They analyzed the multivariate outcomes by multiple Univariate CACE models.

The univariate CACE model was first proposed in last century and became very popular in clinical trails for treatment effect estimation. But limited resources for multivariate CACE analysis exist in the available literature, probably as a result of the intensive computation and the complexity of the analysis.

1.4 Outline

This project mainly focuses on the CACE analysis for multivariate outcomes with an application to the Arthritis Health Journal Study. Results from both the Univariate CACE and the Multivariate CACE analysis are compared in terms of the accuracy of estimates, the type I error rate and the statistical power via a simulation study.

The project is organized as follows: Chapter 2 introduces the background of the Arthritis Health Journal Study and provides descriptive analysis of this motivating example. Chapter 3 describes the technical details of the CACE model, including the discussion about the assumptions and the derivation of likelihood functions and score functions. Chapter 4 shows the simulation study on the comparison of the Univariate CACE and the Multivariate CACE for different sample sizes, effect sizes and variance-covariance matrices. Results from our simulation study suggest that the Multivariate CACE analysis provides better estimates and more powerful tests. Chapter 5 presents our application to the Arthritis Health Journal Study. Chapter 6 offers a brief discussion on findings and limitations and a discussion of future research.

Chapter 2

Motivating Example

2.1 Introduction of the Arthritis Health Journal Study

Arthritis Health Journal is an online tool launched by Dr. Diane Lacaille, a senior scientist at Arthritis Research Centre of Canada. This tool enables rheumatoid arthritis (RA) patients to be actively involved in monitoring their symptoms and disease activity. Patient passports and health journals have been commonly used in chronic diseases to promote active involvement of patients in their care, and have led to better treatment and health outcomes.

During the development stages of this study, both patients and rheumatologists were interviewed. They were asked to provide important insights into the value of an arthritis health journal and how it could be used to improve care. Patients identified the potential benefits of increased self-awareness, better self-management, and improved timing of rheumatologists' visits.

RA needs to be treated early and aggressively to achieve the best long-term health outcomes and prevent bone and joint damage. This online tool is a natural fit for RA patients because it will accommodate early and aggressive treatment and the Treat to Target approach, which involves escalating treatment until the target (little or no inflammation) is met , and modifying treatment when this target is no longer met. If patients can self-monitor their own disease activity, they can provide their health care team with early warnings when targets are not being met, this facilitating the Treat to Target approach.

By using this tool, RA patients are able to asses their disease activity, clearly view results (displayed as remission, low, moderate, or high disease activity), and identify patterns over time. By promoting timeliness of visits to rheumatologists and more accurate information to be shared with rheumatologists, overall disease management should ultimately improve resulting in better outcomes for both RA patients and the health care system. They planed to invite 100 patients with RA to test the Arthritis Health Journal but only 94 patients showed up at baseline.

2.2 Descriptive Analysis

The original dataset consists of the measurements at baseline and at 3 month after the treatment. In order to evaluate the efficacy of this treatment, we focus on the change scores from the baseline; that is, the difference between measurements at 3 month and the measurements at baseline. In the Arthritis Health Journal Study, 94 participants were randomly assigned to the treatment group and the control group at baseline and a total of 45 (47.87%) were allocated to the treatment group. A random sample of 14 participants were selected to take part in face-to-face interviews at 6 month instead of self-reported questionnaires, so their measurements at 3 month were missing and 11 of these 14 participants were in the treatment group. Since they were chosen randomly, we just assume that the missing mechanism is missing completely at random (MCAR) and delete the observations with missing values. Therefore, the final sample size of the Arthritis Health Journal data is 80 in total: 34 in the treatment group and 46 in the control group.

There are a total of 6 health outcomes in the data set and all of them can be treated as continuous variables

- **Effective Consumer 17 Scale:** The overall score of questions about participants and how they manage their disease on a 0 to 100 scale, and 100 indicates "most confident".
- **Manage Symptoms Scale:** The overall score of questions about how they manage their symptoms on a 0 to 10 scale, where 0 indicates "not at all confident" and 10 indicates "totally confident".
- **Manage Disease in General Scale:** The overall score of questions about how they manage their disease in general on a 0 to 10 scale, where 0 indicates "not at all confident" and 10 indicates "totally confident".
- **Communicate with Physician Scale:** The overall score of confidence in communicating with their rheumatologists on a 0 to 10 scale, where 0 indicates "not at all confident" and 10 indicates "totally confident".
- **Partners in Health Scale:** The overall score of their knowledge of disease and treatment on a 0 to 80 scale, where 80 indicates "poor self-management".
- **Satisfaction with Various Aspects of Medical Care:** The overall score of their satisfaction with the content and format of the tool on a 0 to 10 scale, where 0 indicates "completely unsatisfied" and 10 indicates "completely satisfied".

Except for the fifth outcome **Partners in Health Scale**, positive values of change scores for the rest five outcomes represent beneficial treatment effect and negative values represent harmful treatment effect. We first rescale Y_2 to Y_6 to make all outcomes on a 0 to 100 scale for easy comparing.

Additionally, there are a total of 5 binary baseline covariates in the data set

- **Disease Duration**: 0 indicates late disease, that is, having disease for 2 or more years; and 1 indicates early disease (0-2 years).
- **Disease Activity 1/2**: 0 indicates low disease activity (with remission, moderate/low RAPID4 values) and 1 indicates high disease activity (high RAPID4 values).
- **Gender**: 0 indicates female and 1 indicates male.
- **Age**: 0 indicates below the median age (54.50) and 1 otherwise.

There are two other variables in the dataset: one storing the allocation of group assignment and one reflecting the compliance behavior.

- **Group**: 0 indicates the treatment group and 1 indicates the control group; participants in the treatment receive the treatment immediately and for ethical concerns, participants in the control group receive the treatment after 6 months from baseline.
- **High/Low User**: 1 indicates that the number of using the tool is no less than 3 times at 3 month, thus represents high user and 0 represents low user. We define the high users as compliers and low users as non-compliers.

We begin with presenting the descriptive statistics of the change scores for 80 participants. Table 2.1 shows the minimum, median, mean and maximum values of six outcomes grouped by the treatment assignment. An important observation is that Y_1 and Y_5 have a wide range, implying larger variances. The medians for all outcomes are around 0, which confirms the fact that all change scores are centered around 0. It is also worth noticing that the minimum, mean and maximum values for the treatment group are different those for the control group.

Table 2.1: Summary statistics of change scores.

Outcome	Min			Median			Mean			Max		
	A(N = 80)	T(N = 34)	C(N = 46)	A(N = 80)	T(N = 34)	C(N = 46)	A(N = 80)	T(N = 34)	C(N = 46)	A(N = 80)	T(N = 34)	C(N = 46)
Y_1	-30.88	-22.06	-30.88	0.74	0.74	0.74	1.86	2.90	1.09	25.00	25.00	20.59
Y_2	-38.00	-22.00	-38.00	0.00	-1.00	0.00	-0.14	0.57	-0.07	32.00	32.00	26.00
Y_3	-24.00	-18.00	-24.00	2.00	0.00	2.00	1.58	1.29	0.18	46.00	46.00	28.00
Y_4	-40.00	-26.67	-40.00	0.00	0.00	0.00	0.38	3.43	-0.19	60.00	60.00	33.33
Y_5	-43.75	-43.75	-28.75	-0.63	-0.63	-0.63	-1.80	-3.38	-0.50	41.25	23.75	41.25
Y_6	-5.94	-5.94	-4.84	0.00	0.31	0.00	0.37	0.71	0.09	10.16	10.16	6.56

- A indicates all individuals, T indicates the treatment group, C indicates the control group
- Numbers in brackets are sample sizes for different groups

Figure 2.1 shows the distribution of health outcomes for the treatment group, the control group and compliers in the treatment group. We can barely observe the difference between the entire treatment group and the control group. Compliers can not be separated from the non-compliers in the control group, thus we only show the subgroup of compliers in the treatment. For outcome Y_4, Y_5 and Y_6, the subgroup of compliers, as compared

with the entire control group, appears to be in the direction of enjoying the benefit from the treatment. There are no discernable patterns for the other 3 outcomes. Simply comparing the subgroup of compliers with the entire control group can lead to misleading results as the subgroup is not comparable with the control group at the baseline. Therefore, the objective of this project is to develop a method to investigate the complier-average treatment effect combining information from all outcomes.

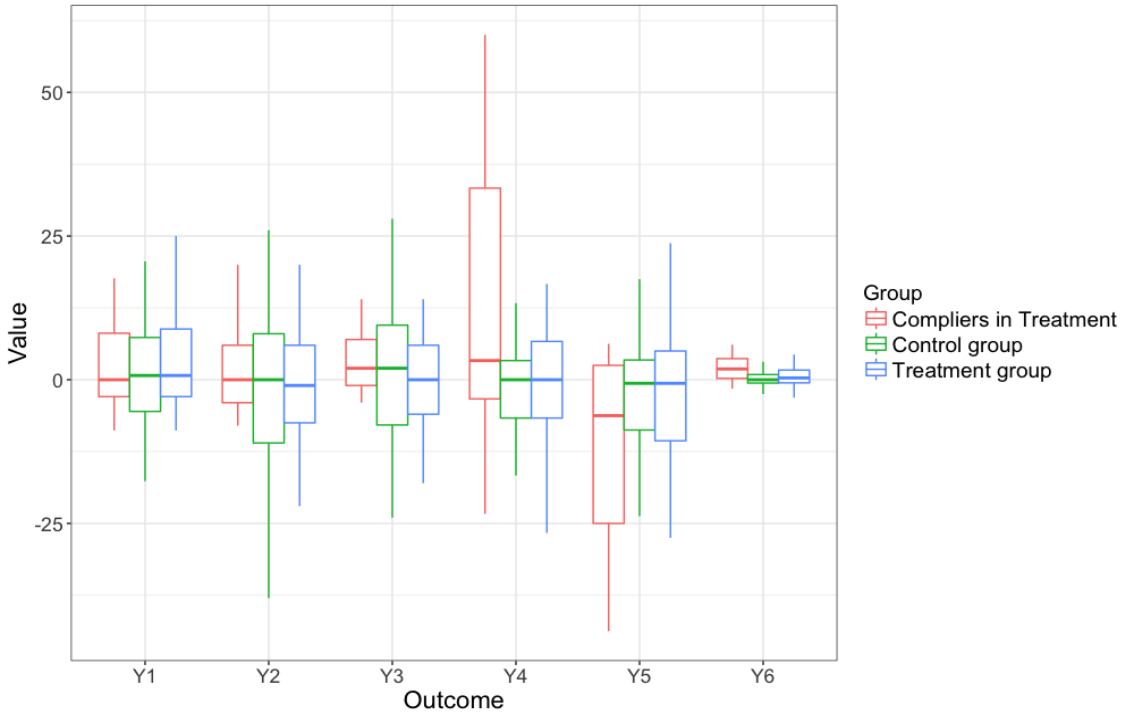

Figure 2.1: Distribution of multivariate outcomes for the treatment group, the control group and the subgroup of compliers: red boxes represent the subgroup of compliers, green represents the control group and blue represents the treatment group.

Chapter 3

Treatment Effect Evaluation Methods

3.1 Notation

To better define the complier-average causal effect (CACE) model, we consider a hypothetical evaluation of the treatment effect of a new treatment **D** on some health outcome **Y** in a population of **N** participants. The initial assignment of participants is stored in the variable **X**:

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ participant is assigned to the treatment,} \\ 0 & \text{if the } i^{th} \text{ participant is assigned to the control.} \end{cases}$$

The actual receipt of the treatment is denoted by the variable **Z**:

$$Z_i = \begin{cases} 1 & \text{if the } i^{th} \text{ participant receives the treatment,} \\ 0 & \text{if the } i^{th} \text{ participant does not receive the treatment.} \end{cases}$$

For clinical trials involving human subjects, the value of the binary variable **Z** is not under investigators' control due to ethical problems. Thus, we write **Z** as a function of **X**. Let $Z_i(X)$ be the indicator of whether the i^{th} participant takes the treatment given assignment X:

$$Z_i(X) = \begin{cases} 1 & \text{if the } i^{th} \text{ participant receives the treatment given assignment } X, \\ 0 & \text{if the } i^{th} \text{ participant does not receive the treatment given assignment } X. \end{cases}$$

In the case of perfect compliance, $Z_i(X) = X_i$ for all participants. Unfortunately, $Z_i(X)$ differs from X_i for various reasons in practice. Similarly, we can define $Y_i(X, Z_i(X))$ as the outcome of the i^{th} participant given the random assignment **X** and the actual receipt **Z**. For univariate analysis, $\mathbf{Y}(X, Z)$ is a vector with N elements, whereas in the multivariate case, $\mathbf{Y}(X, Z)$ is a $N \times k$ matrix (k is the dimension). Define $\underline{Z_i} = (Z_i(0), Z_i(1))$ and $\underline{Y_i} = (Y_i(0, Z_i(0)), Y_i(1, Z_i(1)))$ to be the potential outcomes, which can be partially observed in the experiment.

3.2 Basic Analysis of Treatment Effect

As we mentioned in Chapter 1, intention-to-treat (ITT) analysis and as-treated (AT) analysis are two commonly used approaches for modeling treatment effect. The ITT analysis compares outcomes between participants assigned to the treatment and participants assigned to the control, while the AT analysis compares outcomes of patients who actually received the treatment to outcomes of those who did not. We would like to introduce two assumptions before we define the ITT causal effect and the AT causal effect.

3.2.1 Assumptions for Causal Inference

Assumption 1: SUTVA

Stable Unit Treatment Value Assumption (SUTVA) is a conventional and important limitation in causal effect analysis. SUTVA consists of two components: no interference and well defined potential outcomes. The first component requires that the potential outcome of one unit should be unaffected by the particular assignment of treatment to other units. The second component requires that each participant receives the exactly same version of treatment. SUTVA allows us to write $Y_i(X, Z)$ and $Z_i(X)$ as $Y_i(X_i, Z_i)$ and $Z_i(X_i)$. An experiment would yield biased estimate of causal effect if SUTVA is violated.

Assumption 2: Random Assignment

We also assume the initial treatment assignment X is random; that is, the treatment assignment is independent of all baseline variables. This assumption ensures that each participant has an equal chance of being allocated in any arm. Therefore, observations in the treatment and the control group are exchangeable.

In cases where random assignment would violate ethical standards or in observational studies, the random assignment assumption could be replaced by the Ignorability of Treatment when evaluating the causal effect. The Ignorability assumption simply means that the choice of assignment can be assumed to be effectively random when conditioned on observable characteristics (or baseline variables) of the study objects.

3.2.2 Intention-to-treat Analysis & As-treated Analysis

Under the SUTVA, the causal effect of **X** on **Z** at the individual level can be defined by $Z_i(1) - Z_i(0)$, and accordingly, the causal effect of **X** on **Y** at the unit level can be defined by $Y_i(1, Z_i(1)) - Y_i(0, Z_i(0))$. The causal effect of **X** on **Y** is usually not observable because each participant could be only assigned to either the treatment or the control group; that is, $Y_i(1, Z_i(1))$ and $Y_i(0, Z_i(0))$ are not jointly observable. Since the random assignment eliminates unobserved and confounding factors, we consider the average causal effect.

The average ITT causal effect of **X** on **Z** and the average ITT causal effect of **X** on **Y** are defined as

$$\mathbf{ITT_Z} = \frac{1}{N}\sum_{i=1}^{N}\big(Z_i(1) - Z_i(0)\big), \tag{3.1}$$

and

$$\mathbf{ITT_Y} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1, Z_i(1)) - Y_i(0, Z_i(0))\big). \tag{3.2}$$

The average ITT causal effect can also be calculated by

$$\mathbf{ITT_Y} = \frac{\sum_{i=1}^{N} Y_i X_i}{\sum_{i=1}^{N} X_i} - \frac{\sum_{i=1}^{N} Y_i(1 - X_i)}{\sum_{i=1}^{N}(1 - X_i)}. \tag{3.3}$$

Unlike ITT analysis, the AT analysis only focuses on the actual receipt of treatment and ignores the initial assignment of treatment. Hence, the potential outcome Y only depends on Z, and it can be written as $Y(Z)$. We define the average AT causal effect of **Z** on **Y** as

$$\mathbf{AT_Y} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0))\big), \tag{3.4}$$

which can also be written in the form of (3.3)

$$\mathbf{AT_Y} = \frac{\sum_{i=1}^{N} Y_i Z_i}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N} Y_i(1 - Z_i)}{\sum_{i=1}^{N}(1 - Z_i)}. \tag{3.5}$$

3.3 CACE

3.3.1 Definition of the Compliance Type

There are four types of compliance behavior: compliers, never-takers, always-takers and defiers. Compliers always adhere to the original assignment; never-takers never take the treatment even if they are assigned to the treatment group; always-takers always take the treatment regardless of the initial assignment; and defiers always do the opposite of their assignment. The compliance type takes four possible values:

$$C_i = \begin{cases} c(\text{complier}) & \text{if } Z_i(X) = X, \text{ for } X = 0, 1; \\ n(\text{never-taker}) & \text{if } Z_i(X) = 0, \text{ for } X = 0, 1; \\ a(\text{always-taker}) & \text{if } Z_i(X) = 1, \text{ for } X = 0, 1; \\ d(\text{defier}) & \text{if } Z_i(X) = 1 - X, \text{ for } X = 0, 1. \end{cases}$$

So **C** is a vector with **N** elements and $\mathbf{N_t}$ is the number of participants of type t, where $t \in \{c, n, a, d\}$.

Table 3.1 illustrates the relations among the treatment assignment, the compliance type, the actual receipt of the treatment and the outcomes. This table also shows how Z_i is determined by X_i and C_i.

Table 3.1: Relations among the treatment assignment, the compliance mechanism, the actual receipt of the treatment and the outcomes.

X_i	C_i	$Z_i(X)$	$Y_i(X,Z)$			
			$Y_i(0,0)$	$Y_i(0,1)$	$Y_i(1,0)$	$Y_i(1,1)$
1	c	1	-	-	-	*
0	c	0	*	-	-	-
1	n	0	-	-	*	-
0	n	0	*	-	-	-
1	a	1	-	-	-	*
0	a	1	-	*	-	-
1	d	0	-	-	*	-
0	d	1	-	*	-	-

* represents the observed outcome
- represents the unobserved outcome

3.3.2 Assumptions for CACE

In addition to two assumptions for basic analysis of causal effects, we need three more assumptions to build an identifiable CACE model in this thesis.

Assumption 3: Weak Exclusion Restriction

Weak Exclusion Restriction requires that treatment assignment **X** has no effect on potential outcomes for never-takers and always-takers; that is: for all i, $Z_i(0) = Z_i(1)$, $Y_i(0, Z_i(0)) = Y_i(1, Z_i(1))$. Angrist et al. [1996] proposed Exclusion Restriction for all compliance types:

$$Y_i(X_i, Z_i) = Y_i(X_i', Z_i), \tag{3.6}$$

for all i. This assumption implies that $Y_i(0, Z) = Y_i(1, Z)$ for $\mathbf{Z} = 0, 1$, which means any effect of the assignment **X** on the outcome **Y** must be via the effect of **X** on the actual receipt **Z**. With the support of this assumption, we can write $\mathbf{Y}(X, Z)$ as a function of **Z** alone. This assumption is reasonable in the double-blinded RCTs since participants do not know the treatment assigned to them.

Assumption 4: Strict Monotonicity

This assumption was first proposed by Imbens and Angrist [1994]. Strict Monotonicity restricts the patterns of compliance behavior; that is, for all $i \in \{1, ..., N\}$, $Z_i(1) \geq Z_i(0)$

with inequality for at least one unit. This assumption rules out the occurrence of defiers and requires the presence of at least one complier.

Assumption 5

Since this thesis is motivated by the Arthritis Health Journal Study, we have one more assumption to rule out always-takers in the settings of our CACE model. That is, participants in the control group have no access to the new treatment. Thus, we only have compliers and never-takers in our model.

Aside from the assumptions we have mentioned above, our notation of the assignment $\mathbf{X}$ and the actual receipt $\mathbf{Z}$ also implies that $\mathbf{X}$ and $\mathbf{Z}$ only have two levels: 0 and 1. Hence we do not consider partial compliance. These assumptions hold throughout the rest of this thesis.

3.3.3 CACE

The complier-average causal effect (CACE) is a particular form of the ITT analysis or the AT analysis where inference concerns the average treatment effect within the subgroup of compliers. As defined in (3.2), the ITT effect on $\mathbf{Y}$ can be written as

$$\mathbf{ITT_Y} = \sum_{t\in\{c,n,a,d\}} \frac{N_t\ \mathbf{ITT}_Y^{(t)}}{N}, \tag{3.7}$$

and for $t \in \{c, n, a, d\}$, the ITT effect on $\mathbf{Y}$ for each compliance type can be written as

$$\mathbf{ITT}_\mathbf{Y}^{(\mathbf{t})} = \sum_{\{i|C_i=t\}} \frac{Y_i(1, Z_i(1)) - Y_i(0, Z_i(0))}{N_t}, \tag{3.8}$$

where N_t is the number of participants with compliance type t.

Then, define the CACE of $\mathbf{Z}$ on $\mathbf{Y}$ to be $\mathbf{ITT}_\mathbf{Y}^{(\mathbf{c})}$. For compliers, $Z_i(1) = 1$ and $Z_i(0) = 0$ and based on Assumption 3, Equation (3.8) can be simplified as

$$\mathbf{CACE} = \mathbf{ITT}_\mathbf{Y}^{(\mathbf{c})} = \sum_{\{i|C_i=c\}} \frac{Y_i(1) - Y_i(0)}{N_c}. \tag{3.9}$$

Under Weak Exclusion Restriction assumption, the subgroup of never-takers does not address the causal effect of receiving the new treatment: both $Y_i(1, Z_i(1))$ and $Y_i(0, Z_i(0))$ represent outcomes without taking any treatment. Therefore, we have $\mathbf{ITT}_\mathbf{Y}^{(\mathbf{n})} = 0$.

3.3.4 Likelihood Function

Assume the health outcome $\mathbf{Y}$ is continuous. As defined in (3.9), if we know the exact number of compliers in both treatment and control groups, CACE could be calculated as

$$\mathbf{CACE} = \mathbf{ITT}_{\mathbf{Y}}^{(\mathbf{c})} = \sum_{\{i|C_i=c,X_i=1\}} \frac{Y_i(1)}{N_{c1}} - \sum_{\{j|C_j=c,X_j=0\}} \frac{Y_j(0)}{N_{c2}}, \tag{3.10}$$

where N_{c1} and N_{c2} are the number of compliers in the treatment group and the control group, and $N_{c1} + N_{c2} = N_c$. Unfortunately, the compliance type is unobservable for participants in the control group in many RCTs. For example, in our Arthritis Health Journal Study, all individuals assigned to the control group have no access to the new treatment, so we have no idea what they would do if they were assigned to the treatment group; that is, they could be either compliers or never-takers. In such cases, we could not use (3.10) to get an unbiased estimate of CACE. Instead, we derive the likelihood function and the estimate of CACE is carried out via maximum likelihood.

Univariate Case

Previously, multiple univariate models were used to model multivariate outcomes without considering the correlation among outcomes. So we begin with modeling the univariate outcome and assume the health outcome $\mathbf{Y}$ follows a normal distribution. Without loss of generality, considering the baseline covariates W, we define the probability of a participant being a complier as

$$Pr(C_i = c|W_i = w, \psi) = p_c = \Psi(c, w, \psi), \tag{3.11}$$

and

$$Pr(C_i = n|W_i = w, \psi) = 1 - \Psi(c, w, \psi), \tag{3.12}$$

where we have

$$\Psi(c, w, \psi) = \frac{\exp(w\psi)}{1 + \exp(w\psi)}. \tag{3.13}$$

The compliance rate p_c is unaffected by the initial assignment $\mathbf{X}$ and the actual receipt $\mathbf{Z}$.

Consider those who have been assigned to the treatment group ($X_i = 1$): if $Z_i = 0$, the i^{th} participant is a never-taker, otherwise, he/she is a complier. For those who have been assigned to the control group ($X_i = 0$), we have no information about their compliance behavior. Based on the randomization assumption, there should be the same percent of participants being compliers in two groups. Thus, we assume that the probability of being a complier for participants in the control is p_c, as defined in (3.11).

We further assume that for compliers in the treatment group,

$$\mathbf{Y} \sim \mathbf{N}(\mu_c + \delta_c, \sigma_c^2); \tag{3.14}$$

for compliers in the control group,

$$\mathbf{Y} \sim \mathbf{N}(\mu_c, \sigma_c^2); \tag{3.15}$$

and for non-compliers,

$$\mathbf{Y} \sim \mathbf{N}(\mu_n, \sigma_n^2), \tag{3.16}$$

where $\mathbf{Y}$ is the heath outcome. δ_c is the difference between the compliers in the treatment and the compliers in the control, thus represents the CACE. Since the compliance type is not affected by the treatment assignment and all participants are assigned randomly, there is no difference between non-compliers in different groups. Recall that $\underline{Z_i} = (Z_i(0), Z_i(1))$ and $\underline{Y_i} = (Y_i(0, Z_i(0)), Y_i(1, Z_i(1)))$ are the potential outcomes, which can be partially observed in the experiment. There are three possible patterns of missing and observed data in $(\underline{Z_i}, \underline{Y_i})$ corresponding to the three possible values for $(X_{obs,i}, Z_{obs,i})$: (0, 0), (1, 0), (1, 1), which are displayed in Table 3.2. Define the subsets of units exhibiting each pattern by $S(0,0)$, $S(1,0)$ and $S(1,1)$. For example, for $i \in S(0,0)$, both $Z_i(1)$ and $Y_i(1, Z_i(1))$ are missing. Table 3.3 shows the distribution of $\mathbf{Y}$ under different combinations of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{C}$.

Table 3.2: Possible patterns of missing and observed data.

$(X_{obs,i}, Z_{obs,i})$	$\underline{Z_i}$		$\underline{Y_i}$		Subset
	$Z_i(0)$	$Z_i(1)$	$Y_i(0, Z_i(0))$	$Y_i(1, Z_i(1))$	
(0, 0)	*	-	*	-	$i \in S(0,0)$
(1, 0)	-	*	-	*	$i \in S(1,0)$
(1, 1)	-	*	-	*	$i \in S(1,1)$

* represents the observed outcome
- represents the missing data

Table 3.3: Distribution of $\mathbf{Y}$ under different combinations of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{C}$.

X_i	Z_i	C_i	Y_i
1 (Treatment)	1	c	$N(\mu_c + \delta_c, \sigma_c^2)$
1 (Treatment)	0	n	$N(\mu_n, \sigma_n^2)$
0 (Control)	0	c/n	$p_c N(\mu_c, \sigma_c^2) + (1 - p_c) N(\mu_n, \sigma_n^2)$

Based on Table 3.2 and Table 3.3, we can derive the likelihood function in terms of the observed data separately:

$$\mathcal{L}_i(\theta|X_i^{obs}, Z_i^{obs}, Y_i^{obs}, W_i^{obs}) \propto \begin{cases} \Psi(c,w,\psi)\frac{1}{\sqrt{2\pi\sigma_c^2}}\ \exp\{-\frac{(y_i-(\mu_c+\delta_c))^2}{2\sigma_c^2}\} & \text{if } i \in S(1,1); \\ (1-\Psi(c,w,\psi))\frac{1}{\sqrt{2\pi\sigma_n^2}}\ \exp\{-\frac{(y_i-\mu_n)^2}{2\sigma_n^2}\} & \text{if } i \in S(1,0); \\ \Psi(c,w,\psi)\frac{1}{\sqrt{2\pi\sigma_c^2}}\ \exp\{-\frac{(y_i-\mu_c)^2}{2\sigma_c^2}\} & \\ +(1-\Psi(c,w,\psi))\frac{1}{\sqrt{2\pi\sigma_n^2}}\ \exp\{-\frac{(y_i-\mu_n)^2}{2\sigma_n^2}\} & \text{if } i \in S(0,0), \end{cases} \tag{3.17}$$

where $\theta = (\psi, \mu_c, \mu_n, \delta_c, \sigma_c, \sigma_n)$ is the full parameter vector. The rationale behind the likelihood is that the outcome distribution of never-takers in the control is the same as the outcome distribution of never-takers in the treatment and the control group only consists of compliers and never-takers.

We assume that all compliers have the same variance σ_c^2 regardless of which group they were assign to and similarly, all never-takers have the same variance σ_n^2. The main target of this research is to estimate the method effectiveness and thus we only care about compliers. We want to explore the difference between the compliers in the treatment group and the control group, denoted by δ_c in (3.14).

The actual likelihood function for N individuals is

$$\begin{aligned} & \mathcal{L}(\theta|X^{obs}, Z^{obs}, Y^{obs}, W^{obs}) \\ \propto\ & \prod_{\{i\in S(1,1)\}} \Psi(c,w,\psi)\frac{1}{\sqrt{2\pi\sigma_c^2}}\ \exp\{-\frac{(y_i-(\mu_c+\delta_c))^2}{2\sigma_c^2}\} \\ \times\ & \prod_{\{i\in S(1,0)\}} (1-\Psi(c,w,\psi))\frac{1}{\sqrt{2\pi\sigma_n^2}}\ \exp\{-\frac{(y_i-\mu_n)^2}{2\sigma_n^2}\} \\ \times\ & \prod_{\{i\in S(0,0)\}} \left[\Psi(c,w,\psi)\frac{1}{\sqrt{2\pi\sigma_c^2}}\ \exp\{-\frac{(y_i-\mu_c)^2}{2\sigma_c^2}\} + (1-\Psi(c,w,\psi))\frac{1}{\sqrt{2\pi\sigma_n^2}}\ \exp\{-\frac{(y_i-\mu_n)^2}{2\sigma_n^2}\}\right]. \end{aligned} \tag{3.18}$$

Multivariate Case

Recall that our motivating example consists of 6 health outcomes and we do not want to risk losing any information before we get to the final conclusion. Therefore, we propose a multivariate CACE model to account for correlations among outcomes. The idea is to extend the univariate CACE model to multivariate cases. Naturally, we assume the health outcome $\mathbf{Y}$ follows a multivariate normal distribution.

Similarly, for compliers in the treatment group, we assume

$$\mathbf{Y} \sim \mathbf{MVN}_k(\mu_c + \delta_c,\ \Sigma_c); \tag{3.19}$$

and for compliers in the control group, we assume

$$\mathbf{Y} \sim \mathbf{MVN}_k(\mu_c,\ \Sigma_c); \tag{3.20}$$

and for non-compliers in both groups, we assume

$$\mathbf{Y} \sim \mathbf{MVN}_k(\mu_n,\ \Sigma_n), \tag{3.21}$$

where $\mathbf{MVN}_k$ denotes the k-dimensional normal distribution and $\mathbf{Y}$ is the multivariate health outcome. Unlike the univariate case, μ_c, δ_c, μ_n are vectors and Σ_c, Σ_n are variance-covariance matrices.

Consider the observed likelihood for the i^{th} participant:

$$\mathcal{L}_i(\theta|X_i^{obs}, Z_i^{obs}, Y_i^{obs}, W_i^{obs}) \propto \begin{cases} \Psi(c,w,\psi)\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-(\mu_c+\delta_c))^T\Sigma_c^{-1}(y_i-(\mu_c+\delta_c))\} & \text{if } i \in S(1,1); \\ (1-\Psi(c,w,\psi))\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_n)^T\Sigma_n^{-1}(y_i-\mu_n)\} & \text{if } i \in S(1,0); \\ \Psi(c,w,\psi)\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_c)^T\Sigma_c^{-1}(y_i-\mu_c)\} & \\ +(1-\Psi(c,w,\psi))\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_n)^T\Sigma_n^{-1}(y_i-\mu_n)\} & \text{if } i \in S(0,0), \end{cases} \tag{3.22}$$

where $\theta = (\psi, \mu_c, \mu_n, \delta_c, \Sigma_c, \Sigma_n)$ is the full parameter vector.

Similarly, we assume all compliers have the same variance-covariance matrix Σ_c and all never-takers have the same variance-covariance matrix Σ_n. Modeling k-dimensional outcomes via multiple Univariate CACE models needs to estimate $k-1$ more parameters for p_c and $\frac{1}{2}k(k-1)$ less parameters for covariance per variance-covariance matrix than via the Multivariate CACE model. So Multivariate CACE model needs to estimate $(k-1)^2$ more parameters in total.

The likelihood function for N individuals is then

$$\begin{aligned} & \mathcal{L}(\theta|X^{obs}, Z^{obs}, Y^{obs}, W^{obs}) \\ \propto\ & \textstyle\prod_{\{i\in S(1,1)\}} \Psi(c,w,\psi)\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-(\mu_c+\delta_c))^T\Sigma_c^{-1}(y_i-(\mu_c+\delta_c))\} \\ \times\ & \textstyle\prod_{\{i\in S(1,0)\}} (1-\Psi(c,w,\psi))\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_n)^T\Sigma_n^{-1}(y_i-\mu_n)\} \\ \times\ & \textstyle\prod_{\{i\in S(0,0)\}} \big[\Psi(c,w,\psi)\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_c)^T\Sigma_c^{-1}(y_i-\mu_c)\} \\ +\ & (1-\Psi(c,w,\psi))\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}}\ \exp\{-\frac{1}{2}(y_i-\mu_n)^T\Sigma_n^{-1}(y_i-\mu_n)\}\big]. \end{aligned} \tag{3.23}$$

3.3.5 The Identifiability of the Likelihood Function

A natural question of interest is the identifiability of the derived likelihood function. The identifiable likelihood ensures us to learn the true values of parameters in the CACE model.

Upon a closer inspection on the likelihood function defined in (3.23), the first two parts are identifiable. The third part is the probability density function of a mixed multivariate

normal distribution which might not be identifiable. For example, suppose $\Psi(c, w, \psi) = 0.5$, $\mu_c = \mu_1, \Sigma_c = \Sigma_1$, and $\mu_n = \mu_2, \Sigma_n = \Sigma_2$, and then the mixture part would generate the same value when $\Psi(c, w, \psi) = 0.5$, $\mu_c = \mu_2, \Sigma_c = \Sigma_2$, and $\mu_n = \mu_1, \Sigma_n = \Sigma_1$. Fortunately, the first two parts of the likelihood function yield different values with different sets of parameters. Hence the whole likelihood function is identifiable.

The violation of Weak Exclusion Restriction assumption may lead to non-identifiable CACE model. Without Weak Exclusion Restriction, the potential outcomes for never-takers in the treatment group and the control group are different. In contrast to (3.23), the likelihood function for N individuals should be written as

$$\begin{aligned}
& \mathcal{L}(\theta|X^{obs}, Z^{obs}, Y^{obs}, W^{obs}) \\
\propto\ & \prod_{\{i\in S(1,1)\}} \Psi(c, w, \psi) \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(y_i - (\mu_c + \delta_c))^T \Sigma_c^{-1}(y_i - (\mu_c + \delta_c))\} \\
\times\ & \prod_{\{i\in S(1,0)\}} (1 - \Psi(c, w, \psi)) \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(y_i - (\mu_n + \delta_n))^T \Sigma_n^{-1}(y_i - (\mu_n + \delta_n))\} \\
\times\ & \prod_{\{i\in S(0,0)\}} \Big[\Psi(c, w, \psi) \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(y_i - \mu_c)^T \Sigma_c^{-1}(y_i - \mu_c)\} \\
+\ & (1 - \Psi(c, w, \psi)) \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(y_i - \mu_n)^T \Sigma_n^{-1}(y_i - \mu_n)\}\Big].
\end{aligned} \tag{3.24}$$

Consider a special case when distinct values of θ fail to generate distinct likelihood functions: when $\Psi(c, w, \psi) = 1 - \Psi(c, w, \psi)$ and $N_{\{i\in S(1,1)\}} = N_{\{i\in S(1,0)\}}$, exchanging values of parameters for compliers and never-takers will generate the same likelihood function.

3.3.6 Score Function

In order to speed up the computation process when trying to maximize the log-likelihood function in R, we derive the first derivative with respect to θ. This score function was only used in the simulation study. For simplicity, we write the likelihood function in matrix notation and assume p_c is a constant unaffected by baseline covariates. The variance-covariance matrices Σ_c and Σ_n must be symmetric, positive definite $k \times k$ matrices. Intuitively, only $k(k+1)/2$ parameters are needed to form one variance-covariance matrix. Let η_c and η_n denote the minimal sets of parameters to determine Σ_c and Σ_n; and each of these two sets has $k(k+1)/2$ elements. According to Pinheiro and Bates [1996], the rationale behind the parameterization is to write

$$\Sigma_c = L_c L_c^T \tag{3.25}$$

and

$$\Sigma_n = L_n L_n^T; \tag{3.26}$$

where L_c and L_n are lower triangular matrices. Since there is no constrain of the input values for η_c and η_n, we exponentiate the diagonal elements of L_c and L_n to avoid the occurrence of 0 at diagonal positions of variance-covariance matrices. In order to ensure

the probability of being a complier is between 0 and 1, another transformation is needed before coding the score function. Recall that $\theta = (p_c, \mu_c, \mu_n, \delta_c, \Sigma_c, \Sigma_n)$ is the full parameter vector, assuming dimension k, the elements of θ are $(\theta_1, ..., \theta_{(k+2)^2-3})$. Table 3.4 shows the number of parameters in our Multivariate CACE model.

Table 3.4: The number of parameters in the Multivariate CACE model.

	p_c	μ_c	μ_n	δ_c	Σ_c	Σ_n	Total
# of parameters	1	k	k	k	$\frac{1}{2}k(k+1)$	$\frac{1}{2}k(k+1)$	$(k+2)^2-3$

So p_c is coded as

$$p_c = \frac{\exp(\theta_1)}{1+\exp(\theta_1)}. \tag{3.27}$$

Let $\beta = (\mu_c, \mu_n, \delta_c)^T = (\theta_2, ..., \theta_{3k+1})^T$ be a vector of means, $(\theta_{3k+2}, ..., \theta_{(k-1)(k-4)/2})$ be the lower triangular elements of L_c and $\exp\{(\theta_{((k-1)(k-4)/2)+1}, ..., \theta_{((k-1)(k-4)/2)+k})\}$ be the diagonal elements of L_c. Similarly, we perform the transformation of the rest elements of θ to form L_n.

Another advantage of writing likelihood function in matrix notation is that it would be convenient to add pretreatment covariates if needed for future research. The log-likelihood function in matrix notation is written as

$$\begin{aligned}
& \ell(\theta|X^{obs}, Z^{obs}, Y^{obs}) \\
\propto\ & \Sigma_{\{i\in S(1,1)\}}\{\log\ p_c + \tfrac{1}{2}\log|\Sigma_c|\ + \big(-\tfrac{1}{2}r_{c_i}^T\Sigma_c^{-1}r_{c_i}\big)\} \\
+\ & \Sigma_{\{i\in S(1,0)\}}\{\log(1-p_c) + \tfrac{1}{2}\log|\Sigma_n|\ + \big(-\tfrac{1}{2}r_{n_i}^T\Sigma_n^{-1}r_{n_i}\big)\} \\
+\ & \Sigma_{\{i\in S(0,0)\}}\big[\log\ p_c + \tfrac{1}{2}\log|\Sigma_c|\ + \big(-\tfrac{1}{2}r_{1_i}^T\Sigma_c^{-1}r_{1_i}\big) \\
+\ & \log(1-p_c) + \tfrac{1}{2}\log|\Sigma_n|\ + \big(-\tfrac{1}{2}r_{2_i}^T\Sigma_n^{-1}r_{2_i}\big)\big],
\end{aligned} \tag{3.28}$$

where $r_{c_i} = y_i - W_{c_i}\beta$, $W_{c_i} = \big[\mathbf{I}_{k\times k}\ \mathbf{0}_{k\times k}\ \mathbf{I}_{k\times k}\big]$, $\mathbf{I}_{k\times k}$ is the identity matrix. Accordingly, $r_{n_i} = y_i - W_{n_i}\beta$, $W_{n_i} = \big[\mathbf{0}_{k\times k}\ \mathbf{I}_{k\times k}\ \mathbf{0}_{k\times k}\big]$; $r_{1_i} = y_i - W_{1_i}\beta$, $W_{1_i} = \big[\mathbf{I}_{k\times k}\ \mathbf{0}_{k\times k}\ \mathbf{0}_{k\times k}\big]$ and $r_{2_i} = r_{n_i}$, $W_{2_i} = W_{n_i}$.

Lindstrom and Bates [1988] derived the first derivative of the log-likelihood in matrix form for linear mixed-effects models. With the help of their formula, we can easily derive the score function for our multivariate CACE model. We start from the first element of our parameter vector θ

$$\begin{aligned}
\tfrac{\partial\ell}{\partial\theta_1} &= \Sigma_{\{i\in S(1,1)\}}\tfrac{\exp(\theta_1)}{p_c(1+\exp(\theta_1))^2} \\
&+\Sigma_{\{i\in S(1,0)\}}\tfrac{\exp(\theta_1)}{(p_c-1)(1+\exp(\theta_1))^2} \\
&+\Sigma_{\{i\in S(0,0)\}}\tfrac{L_{1_i}/p_c - L_{2_i}/(1-p_c)}{L_{1_i}+L_{2_i}}\tfrac{\exp(\theta_1)}{(1+\exp(\theta_1))^2},
\end{aligned} \tag{3.29}$$

where

$$L_{1_i} = \frac{p_c}{(2\pi)^{\frac{k}{2}}|\Sigma_c|^{\frac{1}{2}}} \exp\{-\frac{1}{2} r_{1_i}^T \Sigma_c^{-1} r_{1_i}\},$$

and

$$L_{2_i} = \frac{1-p_c}{(2\pi)^{\frac{k}{2}}|\Sigma_n|^{\frac{1}{2}}} \exp\{-\frac{1}{2} r_{2_i}^T \Sigma_n^{-1} r_{2_i}\}.$$

The first derivative with respect to β can be written as

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \Sigma_{\{i \in S(1,1)\}} W_{c_i}^T \Sigma_c^{-1} r_{c_i} \\ &+ \Sigma_{\{i \in S(1,0)\}} W_{n_i}^T \Sigma_n^{-1} r_{n_i} \\ &+ \Sigma_{\{i \in S(0,0)\}} \tfrac{1}{L_{1_i}+L_{2_i}} (L_{1_i} W_{1_i}^T \Sigma_c^{-1} r_{1_i} + L_{2_i} W_{2_i}^T \Sigma_n^{-1} r_{2_i}). \end{aligned} \tag{3.30}$$

Taking the derivative of log-likelihood with respect to $\eta_c = (\theta_{3k+2}, ..., \theta_{(k-1)(k-4)/2+k})$ consists of two parts: the lower triangular elements and the diagonal elements of Σ_c. For θ_j, where $j \in \{3k+2, ..., (k-1)(k-4)/2+k\}$

$$\begin{aligned} \frac{\partial \ell}{\partial \eta_c} &= \Sigma_{\{i \in S(1,1)\}} - \tfrac{1}{2}\big[\mathrm{tr}(\Sigma_c^{-1} \tfrac{\partial \Sigma_c}{\partial \theta_j}) - r_{c_i}^T \Sigma_c^{-1} \tfrac{\partial \Sigma_c}{\partial \theta_j} \Sigma_c^{-1} r_{c_i}\big] \\ &+ \Sigma_{\{i \in S(1,0)\}} 0 \\ &+ \Sigma_{\{i \in S(0,0)\}} \tfrac{1}{L_{1_i}+L_{2_i}} (-\tfrac{1}{2}\big[\mathrm{tr}(\Sigma_c^{-1} \tfrac{\partial \Sigma_c}{\partial \theta_j}) - r_{1_i}^T \Sigma_c^{-1} \tfrac{\partial \Sigma_c}{\partial \theta_j} \Sigma_c^{-1} r_{1_i}\big]), \end{aligned} \tag{3.31}$$

where $\frac{\partial \Sigma_c}{\partial \theta_j} = L_c J^{nm} + J^{mn} L_c^T$ for lower triangular elements and $\frac{\partial \Sigma_c}{\partial \theta_j} = \exp(\theta_j)(L_c J^{nm} + J^{mn} L_c^T)$ for diagonal elements. J^{mn} is the single-entry matrix with 1 at (m, n) and 0 elsewhere, where $n \in \{1, ..., k-1\}$ and $m \in \{n+1, ..., k\}$.

Similarly, taking the derivative with respect to η_n, we have

$$\begin{aligned} \frac{\partial \ell}{\partial \eta_n} &= \Sigma_{\{i \in S(1,1)\}} 0 \\ &+ \Sigma_{\{i \in S(1,0)\}} - \tfrac{1}{2}\big[\mathrm{tr}(\Sigma_n^{-1} \tfrac{\partial \Sigma_n}{\partial \theta_j}) - r_{n_i}^T \Sigma_n^{-1} \tfrac{\partial \Sigma_n}{\partial \theta_j} \Sigma_n^{-1} r_{n_i}\big] \\ &+ \Sigma_{\{i \in S(0,0)\}} \tfrac{1}{L_{1_i}+L_{2_i}} (-\tfrac{1}{2}\big[\mathrm{tr}(\Sigma_n^{-1} \tfrac{\partial \Sigma_n}{\partial \theta_j}) - r_{2_i}^T \Sigma_n^{-1} \tfrac{\partial \Sigma_n}{\partial \theta_j} \Sigma_n^{-1} r_{2_i}\big]), \end{aligned} \tag{3.32}$$

where $\frac{\partial \Sigma_n}{\partial \theta_j} = L_n J^{nm} + J^{mn} L_n^T$ for lower triangular elements and $\frac{\partial \Sigma_n}{\partial \theta_j} = \exp(\theta_j)(L_n J^{nm} + J^{mn} L_n^T)$ for diagonal elements.

3.3.7 Global Likelihood Ratio Test

In order to evaluate the actual treatment effect from a statistical view, we conducted a global likelihood ratio test to explore whether the difference is statistically significant. We tried to test

$$\mathrm{H}_0 : \delta_c = \mathbf{0}_k$$

$$\mathrm{H}_a : \text{The full model is true,}$$

where the full model consists of all parameters and the reduced model sets δ_c to $\mathbf{0}_k$. Consequently, the reduced model is nested within the full model, as required by the likelihood ratio test. The test statistic can be calculated as

$$G = -2(\ell_{reduced}|_{\hat{\theta}_r} - \ell_{full}|_{\hat{\theta}_f}), \tag{3.33}$$

where ℓ is the log-likelihood, $\hat{\theta}_r$ is the MLE obtained from the reduced model and $\hat{\theta}_f$ is the MLE obtained from the full model. The distribution of the test statistic is approximately an chi-square distribution with degree of freedom k (the dimension of the outcomes).

For multiple Univariate CACE models, the significance cut-off value should be α/k according to the Bonferroni correction, where α is the desired overall alpha level. If any of these k hypotheses is rejected, the treatment effect is considered significant.

3.3.8 Parametric Bootstrap Test

We have found that the mixture part of our CACE model, the sample size and the compliance rate have an influence on the convergence of G. Our simulation results suggest that the type I error inflates under a low compliance rate for moderate sample sizes. If we get rid of the mixture part of the Multivariate CACE model, the estimated type I error drops a lot. We could also control the type I error by improving the compliance or increasing the sample size. But in the case of a low compliance rate and limited sample size, an alternative approach needs to be considered to ensure the accuracy of our test.

Parametric bootstrap turns out to be a good alternative. Bootstrapping is commonly used to estimate the sampling distribution by drawing samples from the estimated population with replacement. The following steps describe how we performed the parametric bootstrap test on each simulated dataset

- Maximized $\ell_{reduced}$ and ℓ_{full} to get $\hat{\theta}_r$ and $\hat{\theta}_f$, and calculated the LR test statistic G_0.
- Generated 100 new datasets using $\hat{\theta}_r$ as the true values of parameters.
- Calculated the new LR test statistic G_{B_i} for the i^{th} dataset, where $i \in \{1, 2, ..., 100\}$.
- Sorted $\{G_{B_1}, ..., G_{B_{100}}\}$ in increasing order to get $\{G_{B_1^*}, ..., G_{B_{100}^*}\}$.
- If $G_0 > G_{B_{95}^*}$, reject H_0.

In this way, we managed to get the estimated distribution of the test statistic (G) when the asymptotic distribution of G is inaccurate for a finite sample size. We expect to observe a clear improvement in estimated type I error by performing the parametric bootstrap test. The results will be displayed in Chapter 4 and Chapter 5.

Chapter 4

Simulation Study

We conducted a simulation study to show the advantages of the proposed Multivariate CACE model when health outcomes are multivariate.

4.1 Design of Study

Under the assumption that the health outcome $\mathbf{Y}$ follows a multivariate normal distribution, the outcomes of the compliers in both groups and the outcomes of never-takers were generated from $\mathbf{MVN}_k(\mu_c + \delta_c, \Sigma_c)$, $\mathbf{MVN}_k(\mu_c, \Sigma_c)$ and $\mathbf{MVN}_k(\mu_n, \Sigma_n)$, where $\mu_c = (1, 1, 1, 2, 2, 2)$ and $\mu_n = (2, 2, 2, 1, 1, 1)$. The effect size δ_c is a vector with same elements. Nine values ranging from 0 to 5 were selected as the effect size: $(0, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5)$, and we set sample size to 100, 200, 500 for non-zero effect sizes and an additional 1000 when effect size equals 0. To keep consistent with our real data example, we set dimension $k = 6$ for all scenarios. For simplicity, we assume the compliance rate is unaffected by baseline covariates and set p_c to 0.4, which is close to the MLE of p_c from our real data example. Given sample size N, the group allocation of participants was stored in vector $\mathbf{X}$ with first $N/2$ participants assigned to the treatment group and the rest assigned to the control group. As for the actual treatment receipt $\mathbf{Z}$, set Z_i to 0 for all units in the control group. For the treatment group, set Z_i to 1 with probability p_c.

In our simulation study, the statistical power of global LR test was used to evaluate the performance of both CACE models, as shown in Figure 4.3 . The power is affected by the sample size, the effect size and variance, thus we considered different scenarios by varying these three factors. For scientific meaning, we assumed the variance-covariance matrices were of the same magnitude. Two different sets of values were considered for variance-covariance matrices, which are presented as follows:

$$\Sigma_{c_1} = \begin{bmatrix} 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 2 & -1 & 0 & 1 & -2 \\ -1 & -1 & 2 & 0 & -1 & 0 \\ -1 & 0 & 0 & 7 & 1 & 1 \\ 1 & 1 & -1 & 1 & 6 & -1 \\ -1 & -2 & 0 & 1 & -1 & 7 \end{bmatrix}, \tag{4.1}$$

$$\Sigma_{n_1} = \begin{bmatrix} 1 & 1 & -1 & -1 & -1 & 1 \\ 1 & 2 & -2 & -2 & -2 & 2 \\ -1 & -2 & 3 & 1 & 1 & -3 \\ -1 & -2 & 1 & 7 & 5 & -3 \\ -1 & -2 & 1 & 5 & 8 & -4 \\ 1 & 2 & -3 & -3 & -4 & 9, \end{bmatrix}; \tag{4.2}$$

and

$$\Sigma_{c_2} = \begin{bmatrix} 9 & 3 & 3 & -3 & -3 & 6 \\ 3 & 10 & -2 & -7 & -7 & 8 \\ 3 & -2 & 11 & 7 & -2 & 6 \\ -3 & -7 & 7 & 18 & -3 & 1 \\ -3 & -7 & -2 & 3 & 19 & -16 \\ 6 & 8 & 6 & 1 & -16 & 26 \end{bmatrix}, \tag{4.3}$$

$$\Sigma_{n_2} = \begin{bmatrix} 9 & 3 & 3 & -3 & -3 & -3 \\ 3 & 10 & -2 & 2 & -2 & 2 \\ 3 & -2 & 11 & 1 & -1 & -5 \\ -3 & 2 & 1 & 12 & 0 & -2 \\ 3 & -2 & -1 & 0 & 13 & 1 \\ -3 & 2 & -5 & -2 & 1 & 14, \end{bmatrix}. \tag{4.4}$$

The corresponding correlation matrices are as follows

$$\varrho_{c_1} = \begin{bmatrix} 1 & 0.71 & -0.71 & -0.38 & 0.41 & -0.38 \\ 0.71 & 1 & -0.50 & 0 & 0.29 & -0.53 \\ -0.71 & -0.50 & 1 & 0 & -0.29 & 0 \\ -0.38 & 0 & 0 & 1 & 0.15 & 0.14 \\ 0.41 & 0.29 & -0.29 & 0.15 & 1 & -0.15 \\ -0.38 & 0.53 & 0 & 0.14 & -0.15 & 1, \end{bmatrix}; \tag{4.5}$$

$$\varrho_{n_1} = \begin{bmatrix} 1 & 0.71 & -0.58 & -0.38 & -0.35 & 0.33 \\ 0.71 & 1 & -0.82 & -0.53 & -0.50 & 0.47 \\ -0.58 & -0.82 & 1 & 0.22 & 0.20 & -0.58 \\ -0.38 & -0.53 & 0.22 & 1 & 0.67 & -0.38 \\ -0.35 & -0.50 & 0.20 & 0.67 & 1 & -0.47 \\ 0.33 & 0.47 & -0.58 & -0.38 & -0.47 & 1, \end{bmatrix}. \tag{4.6}$$

$$\varrho_{c_2} = \begin{bmatrix} 1 & 0.32 & 0.30 & -0.24 & -0.23 & 0.39 \\ 0.32 & 1 & -0.19 & -0.52 & -0.51 & 0.50 \\ 0.30 & -0.19 & 1 & 0.50 & -0.14 & 0.35 \\ -0.24 & -0.52 & 0.50 & 1 & -0.16 & 0.05 \\ -0.23 & -0.51 & -0.14 & -0.16 & 1 & -0.72 \\ 0.39 & 0.50 & 0.35 & 0.05 & -0.72 & 1, \end{bmatrix}; \tag{4.7}$$

$$\varrho_{n_2} = \begin{bmatrix} 1 & 0.32 & 0.30 & -0.29 & 0.28 & -0.27 \\ 0.32 & 1 & -0.19 & 0.18 & -0.18 & 0.17 \\ 0.30 & -0.19 & 1 & 0.09 & -0.08 & -0.40 \\ -0.29 & 0.18 & 0.09 & 1 & 0 & -0.15 \\ 0.28 & -0.18 & -0.08 & 0 & 1 & 0.07 \\ -0.27 & 0.17 & -0.40 & -0.15 & 0.07 & 1, \end{bmatrix}. \tag{4.8}$$

4.2 Results

In this section, we will only focus on the estimates of effect size δ_c and p_c, and the difference in statistical power of multivariate and multiple univariate tests.

4.2.1 Point Estimate

The maximum likelihood estimates (MLEs) of δ_c shown in Table 4.1 were obtained from 500 simulated datasets. Since the effect size has little influence on the accuracy of the estimates, we only selected three values of δ_c. Figure 4.1 presents the distribution of MLEs for different scenarios and each panel shows the comparison of two CACE models. Panels in the same row share the same effect size and panels in the same column share the same sample size. The selected sample sizes are $100, 200, 500$ and the chosen effect sizes are $0, 0.5, 1$. The effect of the variance of outcomes on the distribution of the estimates can be observed from Figure 4.1 because Y_1,..., Y_6 have different variances. Therefore, we only show the results for the smaller variance-covariance matrices (Σ_{c_1}, Σ_{n_1}).

By inspection of Table 4.1 and Figure 4.1, we observe that:

(a) As expected, the finite sample bias for the maximum likelihood estimates (MLEs) of means decreases as sample size increases for both Multivariate CACE and Univariate

Table 4.1: Maximum likelihood estimates of δ_c for the Multivariate CACE and the multiple Univariate CACE.

Σ	δ_c	N	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	0	100	-0.010 (-0.043)	0.009 (-0.038)	-0.027 (-0.133)	-0.019 (0.052)	-0.044 (-0.025)	0.005 (0.164)
		200	0.008 (-0.005)	-0.007 (-0.074)	-0.017 (-0.046)	-0.019 (0.059)	0.044 (0.086)	0.014 (0.052)
		500	-0.002 (-0.002)	-0.001 (-0.012)	-0.013 (-0.038)	0.012 (0.036)	-0.032 (-0.011)	0.009 (0.028)
	0.5	100	0.473 (0.424)	0.497 (0.355)	0.473 (0.449)	0.608 (0.715)	0.468 (0.531)	0.574 (0.589)
		200	0.497 (0.497)	0.484 (0.454)	0.475 (0.461)	0.510 (0.521)	0.475 (0.490)	0.551 (0.599)
		500	0.495 (0.489)	0.494 (0.483)	0.502 (0.504)	0.503 (0.535)	0.475 (0.469)	0.505 (0.528)
	1	100	0.990 (0.905)	1.024 (0.904)	0.993 (0.936)	1.067 (1.095)	1.032 (1.085)	0.971 (0.982)
		200	1.004 (0.992)	1.002 (0.961)	1.009 (0.976)	0.974 (1.039)	1.024 (1.051)	0.906 (0.945)
		500	1.003 (1.004)	0.100 (0.989)	0.995 (1.001)	0.998 (0.996)	0.980 (1.005)	0.981 (0.974)
2	0	100	0.012 (-0.076)	0.080 (-0.098)	-0.001 (0.009)	0.021 (0.083)	-0.143 (0.027)	0.115 (0.120)
		200	0.009 (-0.019)	0.042 (-0.090)	0.031 (-0.048)	0.013 (-0.048)	-0.084 (-0.029)	0.074 (0.061)
		500	-0.037 (-0.057)	0.0003 (-0.011)	-0.028 (-0.031)	-0.001 (0.035)	-0.030 (-0.024)	-0.005 (-0.073)
	0.5	100	0.395 (0.317)	0.482 (0.377)	0.440 (0.278)	0.582 (0.656)	0.524 (0.618)	0.410 (0.447)
		200	0.495 (0.317)	0.50 (0.450)	0.485 (0.393)	0.508 (0.495)	0.518 (0.492)	0.419 (0.304)
		500	0.495 (0.469)	0.507 (0.469)	0.512 (0.507)	0.503 (0.490)	0.474 (0.464)	0.522 (0.489)
	1	100	0.972 (0.857)	1.002 (0.870)	1.036 (0.859)	1.052 (1.113)	0.944 (1.056)	1.077 (0.973)
		200	1.071 (0.979)	1.039 (0.939)	0.983 (0.896)	0.921 (0.934)	1.027 (1.049)	0.971 (0.973)
		500	1.032 (1.037)	1.016 (0.997)	1.024 (1.019)	1.033 (1.030)	0.961 (0.997)	1.021 (1.049)
	2	100	2.023 (1.928)	2.105 (1.966)	1.974 (1.973)	1.990 (2.070)	1.905 (1.879)	2.088 (2.032)
		200	1.968 (1.923)	1.962 (1.922)	1.973 (1.946)	2.026 (2.029)	2.006 (1.928)	1.968 (1.957)
		500	1.987 (1.928)	1.996 (1.964)	1.960 (1.918)	1.9790 (1.988)	2.020 (2.002)	1.916 (1.901)

- N indicates the size of the simulated population.
- δ_c indicates the true value of the effect size : we assume the same effect size for all dimensions.
- Σ indicates different variance-covariance matrices: 1 represents the smaller variance-covariance matrices ($\Sigma_{c_1}, \Sigma_{n_1}$) and 2 represents the larger variance-covariance matrices ($\Sigma_{c_2}, \Sigma_{n_2}$).
- Values in brackets are the mean MLEs over 500 datasets for multiple Univariate CACE models.

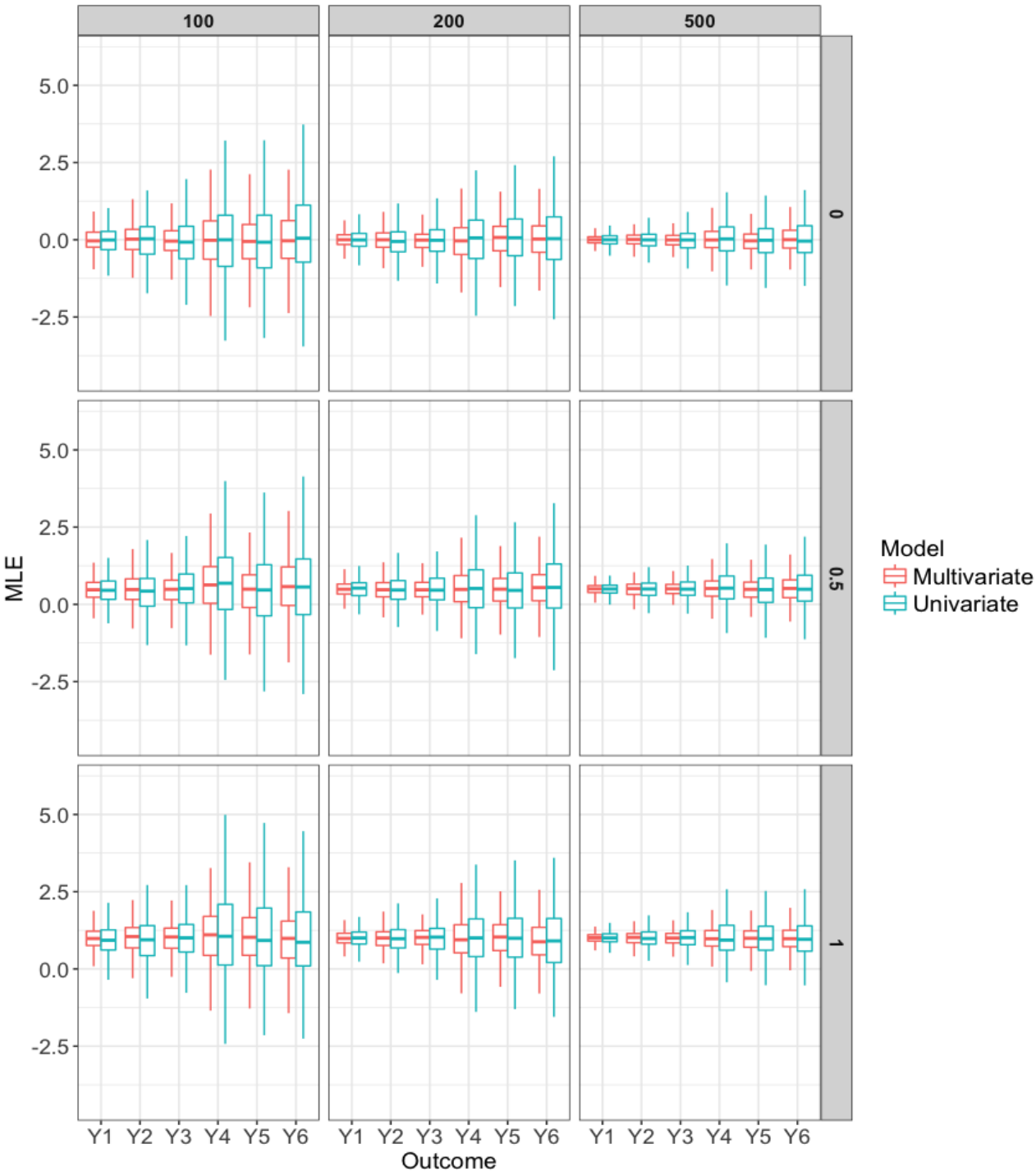

Figure 4.1: Comparison of the distribution of MLEs for two CACE models with smaller variance-covariance matrices ($\Sigma_{c_1}, \Sigma_{n_1}$): red boxes represent the results for the Multivariate CACE and green boxes represent the results for the multiple Univariate CACE.

CACE models and the distribution of MLEs is unaffected by the magnitude of effect size. As defined in (4.1), the variances of (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) are $(1, 2, 2, 7, 6, 7)$,

which could be the proper explanation for different variances of the estimates for different outcomes.

(b) It is worth noting that the variation of estimates is always bigger and the estimates are less accurate for multiple Univariate CACE models in every scenario. A possible reason is that Univariate CACE models the multivariate health outcomes separately and ignores the correlations among outcomes. In conclusion, our proposed Multivariate CACE model outperforms multiple Univariate CACE models in precise estimation.

Table 4.2: Estimates of compliance rate for the Multivariate CACE and the Univariate CACE under different sample sizes.

Sample Size	p_c	Univariate CACE							Multivariate CACE
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Mean	
100	0.4	0.397	0.398	0.395	0.396	0.397	0.396	0.397	0.395
200	0.4	0.399	0.399	0.400	0.398	0.398	0.399	0.399	0.399
500	0.4	0.399	0.400	0.400	0.399	0.399	0.400	0.399	0.399

- Effect size $\delta_c = 0$

The estimates of compliance rate are displayed in Table 4.2, from which we can observe that the mean estimates of p_c from univariate models are very similar to the estimates from the multivariate model.

4.2.2 Interval Estimate

According to the asymptotic normality of MLE, we have

$$\hat{\delta}_c \to \mathbf{MVN}_k(\delta_c, I_N(\delta_c)^{-1}), \tag{4.9}$$

where $I_N(\delta_c)$ is the Fisher information for N individuals. In our simulation study, $I_N(\delta_c)$ is estimated by evaluating the negative Hessian matrix at $\hat{\delta}_c$.

Confidence regions are multivariate extensions of univariate confidence intervals, which should be a k-dimension ellipsoid centered at $\hat{\delta}_c$. Since it is hard to imagine the shape of a high dimensional ellipsoid, simultaneous confidence intervals are more useful in practice. Simultaneous confidence intervals for vector δ_c require that a group of confidence intervals all include true values of δ_c at some confidence level. Simultaneous confidence intervals concentrates on linear scalar functions of δ_c, of the form $t'\delta_c$ for any $t \in R^6$. According to delta method, $t'\hat{\delta}_c$ asymptoticly follows $\mathbf{N}(t'\delta_c, t'I_N(\delta_c)^{-1}t)$ as $N \to \infty$. We managed to calculate the confidence intervals for single element of δ_c by setting a set of vectors $t_1, ..., t_6$, where $t_1 = (1, 0, ..., 0)^T, t_2 = (0, 1, 0, ..., 0)^T$, ..., and $t_6 = (0, ..., 0, 1)^T$, thus the confidence

interval for each element of δ_c can be calculated as

$$t'\hat{\delta}_c \pm c\sqrt{t' I_N(\hat{\delta}_c)^{-1} t}, \tag{4.10}$$

where c is the appropriate critical value. Simultaneous confidence intervals is based on both the individual confidence level and the number of confidence intervals. We used the Bonferroni correction to limit the probability that one or more of the confidence intervals does not contain the true value to a maximum of α.

Table 4.3 shows the mean 95% simultaneous confidence intervals (CIs) based on 500 datasets. Figure 4.2 presents the distribution of the length of the confidence intervals. Panels in the same row share the same effect size and panels in the same column share the same sample size.

By inspection of Table 4.3 and Figure 4.2 , we observe that

(a) Overall, as sample size increases and variance decreases, the width of CIs decreases correspondingly. The effect size is not related to the length of CIs.

(b) While the Bonferroni correction was applied to both the Multivariate CACE and the Univariate CACE, the multivariate model has shorter confidence intervals. Furthermore, multivariate analysis has smaller variance of the length of the confidence intervals.

Additionally, we calculated the coverage rate for confidence intervals. The coverage rate is defined as the proportion of times that the confidence intervals contain the true δ_c values. We have found that the true coverage rate is always less than the nominal coverage probability, which is set to 0.95 in our simulation study. We are happy to observe higher coverage rate of confidence intervals produced by the Multivariate CACE from Table 4.4. Coverage rate increases as sample size increases and remains unaffected by the effect size and the magnitude of variance-covariance matrices.

Table 4.3: 95% simultaneous confidence intervals of δ_c.

Σ	δ_c	N	Model	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	0	100	Mul	(-0.854, 0.833)	(-1.188, 1.206)	(-1.224, 1.169)	(-2.254, 2.215)	(-2.128, 2.040)	(-2.240, 2.250)
			Uni	(-1.174, 1.089)	(-1.641, 1.566)	(-1.829, 1.564)	(-3.043, 3.147)	(-2.986, 2.937)	(-3.108, 3.435)
		200	Mul	(-0.602, 0.617)	(-0.872, 0.859)	(-0.889, 0.856)	(-1.621, 1.584)	(-1.447, 1.535)	(-1.600, 1.629)
			Uni	(-0.830, 0.819)	(-1.290, 1.142)	(-1.333, 1.241)	(-2.263, 2.381)	(-2.165, 2.336)	(-2.393, 2.498)
		500	Mul	(-0.389, 0.384)	(-0.552, 0.551)	(-0.564, 0.539)	(-1.004, 1.031)	(-0.980, 0.915)	(-1.019, 1.039)
			Uni	(-0.520, 0.516)	(-0.782, 0.759)	(-0.826, 0.786)	(-1.483, 1.555)	(-1.494, 1.472)	(-1.545, 1.602)
	0.5	100	Mul	(-0.375, 1.321)	(-0.709, 1.702)	(-0.735, 1.680)	(-1.629, 2.845)	(-1.631, 2.567)	(-1.710, 2.858)
			Uni	(-0.726, 1.575)	(-1.297, 2.006)	(-1.268, 2.167)	(-2.468, 3.897)	(-2.488, 3.551)	(-2.692, 3.870)
		200	Mul	(-0.108, 1.102)	(-0.380, 1.348)	(-0.392, 1.341)	(-1.086, 2.105)	(-1.014, 1.964)	(-1.064, 2.165)
			Uni	(-0.320, 1.313)	(-0.768, 1.676)	(-0.812, 1.735)	(-1.813, 2.854)	(-1.780, 2.760)	(-1.829, 3.027)
		500	Mul	(0.109, 0.880)	(-0.058, 1.046)	(-0.050, 1.053)	(-0.515, 1.521)	(-0.476, 1.426)	(-0.528, 1.539)
			Uni	(-0.031, 1.009)	(-0.291, 1.257)	(-0.316, 1.325)	(-0.986, 2.056)	(-1.001, 1.938)	(-1.054, 2.110)
	1	100	Mul	(0.140, 1.841)	(-0.185, 2.233)	(-0.228, 2.213)	(-1.156, 3.290)	(-1.066, 3.129)	(-1.290, 3.231)
			Uni	(-0.227, 2.037)	(-0.734, 2.543)	(-0.801, 2.673)	(-2.024, 4.215)	(-1.903, 4.072)	(-2.259, 4.224)
		200	Mul	(0.393, 1.615)	(0.130, 1.874)	(0.134, 1.884)	(-0.632, 2.580)	(-0.478, 2.527)	(-0.728, 2.540)
			Uni	(0.160, 1.825)	(-0.266, 2.188)	(-0.306, 2.258)	(-1.322, 3.399)	(-1.243, 3.344)	(-1.484, 3.374)
		500	Mul	(0.616, 1.390)	(0.447, 1.552)	(0.443, 1.547)	(-0.019, 2.014)	(0.029, 1.930)	(-0.051, 2.013)
			Uni	(0.481, 1.526)	(0.216, 1.763)	(0.170, 1.832)	(-0.513, 2.505)	(-0.467, 2.477)	(-0.618, 2.566)
2	0	100	Mul	(-2.612, 2.636)	(-2.671, 2.831)	(-2.903, 2.901)	(-3.676, 3.718)	(-3.912, 3.625)	(-4.224, 4.454)
			Uni	(-3.598, 3.446)	(-3.869, 3.672)	(-3.939, 3.958)	(-4.581, 4.747)	(-4.816, 4.870)	(-5.266, 5.506)
		200	Mul	(-1.866, 1.884)	(-1.925, 2.009)	(-2.026, 2.088)	(-2.610, 2.637)	(-2.758, 2.590)	(-3.007, 3.154)
			Uni	(-2.631, 2.593)	(-2.771, 2.769)	(-2.952, 2.856)	(-3.413, 3.317)	(-3.513, 3.455)	(-3.748, 3.870)
		500	Mul	(-1.241, 1.167)	(-1.253, 1.259)	(-1.344, 1.287)	(-1.671, 1.668)	(-1.731, 1.672)	(-1.976, 1.963)
			Uni	(-1.771, 1.657)	(-1.823, 1.801)	(-1.943, 1.880)	(-2.099, 2.169)	(-2.238, 2.191)	(-2.471, 2.325)
	0.5	100	Mul	(-2.262, 3.051)	(-2.299, 3.262)	(-2.448, 3.328)	(-3.096, 4.259)	(-3.283, 4.331)	(-3.969, 4.788)
			Uni	(-3.627, 3.900)	(-3.370, 4.125)	(-3.687, 4.244)	(-4.082, 5.394)	(-4.211, 5.447)	(-4.918, 5.812)
		200	Mul	(-1.483, 2.357)	(-1.493, 2.474)	(-1.597, 2.567)	(-2.129, 3.145)	(-2.176, 3.213)	(-2.777, 3.460)
			Uni	(-2.383, 3.016)	(-2.371, 3.271)	(-2.552, 3.337)	(-2.933, 3.922)	(-3.046, 4.031)	(-3.561, 4.168)
		500	Mul	(-0.709, 1.699)	(-0.740, 1.753)	(-0.802, 1.825)	(-1.161, 2.166)	(-1.224, 2.173)	(-1.440, 2.484)
			Uni	(-1.242, 2.180)	(-1.327, 2.264)	(-1.384, 2.397)	(-1.640, 2.620)	(-1.746, 2.674)	(-1.893, 2.870)
	1	100	Mul	(-1.679, 3.622)	(-1.750, 3.755)	(-1.857, 3.928)	(-2.618, 4.721)	(-2.823, 4.711)	(-3.298, 5.452)
			Uni	(-2.719, 4.433)	(-2.904, 4.645)	(-3.108, 4.826)	(-3.485, 5.711)	(-3.689, 5.800)	(-4.286, 6.399)
		200	Mul	(-0.814, 2.955)	(-0.928, 3.006)	(-1.071, 3.037)	(-1.628, 3.524)	(-1.641, 3.695)	(-2.120, 4.061)
			Uni	(-1.640, 3.597)	(-1.827, 3.705)	(-2.003, 3.975)	(-2.463, 4.331)	(-2.442, 4.541)	(-2.865, 4.811)
		500	Mul	(-0.171, 2.235)	(-0.238, 2.270)	(-0.291, 2.339)	(-0.632, 2.698)	(-0.746, 2.667)	(-0.941, 2.982)
			Uni	(-0.681, 2.755)	(-0.822, 2.815)	(-0.889, 2.926)	(-1.118, 3.177)	(-1.228, 3.223)	(-1.361, 3.459)
	2	100	Mul	(-0.640, 4.687)	(-0.639, 4.848)	(-0.911, 4.860)	(-1.671, 5.651)	(-1.865, 5.674)	(-2.282, 6.459)
			Uni	(-2.719, 4.433)	(-2.904, 4.645)	(-3.108, 4.826)	(-3.485, 5.711)	(-3.689, 5.800)	(-4.286, 6.399)
		200	Mul	(0.058, 3.877)	(-0.015, 3.940)	(-0.100, 4.045)	(-0.590, 4.643)	(-0.674, 4.686)	(-1.117, 5.053)
			Uni	(-0.753, 4.598)	(-0.913, 4.756)	(-1.007, 4.898)	(-1.317, 5.375)	(-1.547, 5.404)	(-1.826, 5.776)
		500	Mul	(0.783, 3.191)	(0.742, 3.251)	(0.640, 3.280)	(0.312, 3.646)	(0.320, 3.720)	(-0.051, 3.884)
			Uni	(0.197, 3.659)	(0.146, 3.782)	(0.007, 3.829)	(-0.165, 4.141)	(-0.224, 4.229)	(-0.502, 4.305)

- N indicates the size of the simulated population.
- δ_c indicates the true value of the effect size : we assume the same effect size for all dimensions.
- Σ indicates different variance-covariance matrices: 1 represents the smaller variance-covariance $(\Sigma_{c_1}, \Sigma_{n_1})$ matrices and 2 represents the larger variance-covariance matrices $(\Sigma_{c_2}, \Sigma_{n_2})$.

4.2.3 Comparison of Statistical Power

As mentioned in the last Chapter, we performed a global likelihood ratio test to evaluate the significance of treatment effect. A natural question of interest is whether the global likelihood ratio test is a powerful test. The statistical power is used to compare the per-

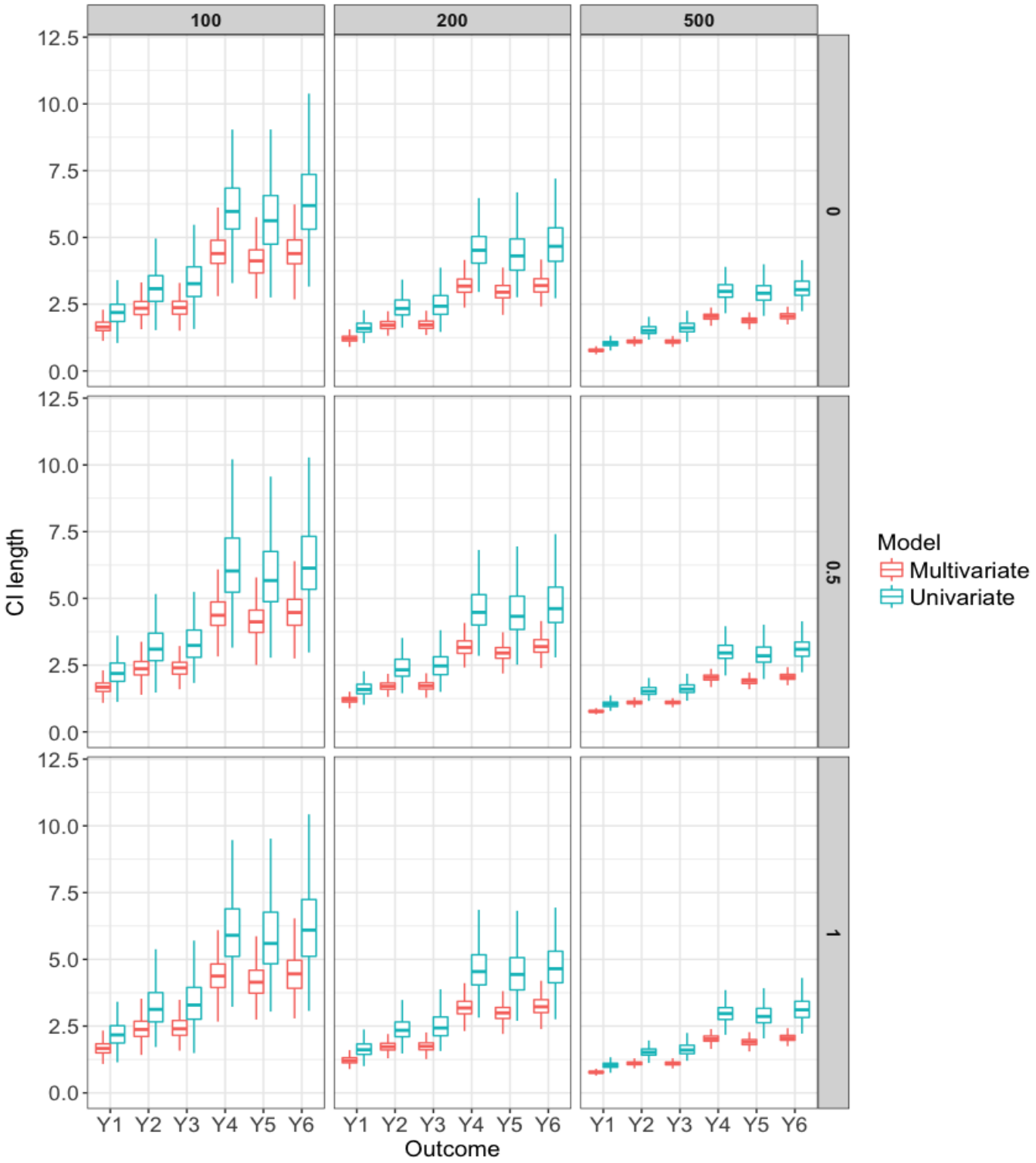

Figure 4.2: Distribution of the length of confidence intervals: red boxes represent the results for the Multivariate CACE and green boxes represent the results for the multiple Univariate CACE.

Table 4.4: Coverage rate of confidence intervals.

Σ	δ_c	N	Multivariate CACE	Univariate CACE
1	0	100	0.890	0.762
		200	0.906	0.866
		500	0.946	0.936
	0.5	100	0.896	0.748
		200	0.934	0.868
		500	0.936	0.932
	1	100	0.882	0.742
		200	0.930	0.864
		500	0.938	0.926
2	0	100	0.886	0.816
		200	0.938	0.876
		500	0.930	0.926
	0.5	100	0.880	0.808
		200	0.940	0.874
		500	0.950	0.920
	1	100	0.884	0.822
		200	0.942	0.868
		500	0.946	0.920
	2	100	0.884	0.814
		200	0.934	0.872
		500	0.934	0.924

- N indicates the size of the simulated population.
- δ_c indicates the true value of the effect size : we assume the same effect size for all dimensions.
- Σ indicates different variance-covariance matrices: 1 represents the smaller variance-covariance matrices $(\Sigma_{c_1}, \Sigma_{n_1})$ and 2 represents the larger variance-covariance matrices $(\Sigma_{c_2}, \Sigma_{n_2})$.

formance of the multivariate model and multiple univariate models. All power analysis is done at the same significance level ($\alpha = 0.05$).

We generated 500 datasets for each setting and for each simulated dataset, we tried to test

$$\mathrm{H}_0 : \delta_c = \mathbf{0}_6.$$

The power of this test can be estimated as

$$\text{power} = Pr(\text{reject } \mathrm{H}_0 | \mathrm{H}_a \text{is true}) = \frac{N_{rej}}{500}, \tag{4.11}$$

where N_{rej} is the number of tests that have been correctly rejected.

Figure 4.3 presents the power curves produced by the Multivariate CACE, the Univariate CACE and parametric bootstrap tests. Panels in the same column represent the results

for the same sample size. Each panel shows the relationship between the power and the effect size under fixed variance and sample size. We observe from the plot that

(a) Overall, the power increases sharply as the effect size increases and the multivariate model has the steepest slopes. The power can be improved by at most 0.7 by performing the multivariate test rather than multiple univariate tests for moderate sample size when the variance is small. When focusing on panels in the same column, we notice that the larger variance leads to lower power under the same effect size. The difference between two CACE models decreases as sample size grows.

(b) We start from the first row, when $N = 100$, the power of the multivariate test reaches 1 after the effect size grows over 1. As the sample size increases to 200 and 500, we could observe a sharp increase of power within small changes in the effect size and the power reaches 1 earlier.

(c) For the larger variance-covariance matrices, we observe a similar trend of power except when the effect sizes are comparably small. In cases where the true effect sizes are too small, compared to the corresponding variance, and the sample size is not large enough, we could observe some little fluctuations for CACE models by chance.

(d) The parametric bootstrap test performs quite robust for different values of the variance-covariance matrices and the sample size. It performs much better than the Univariate CACE and slightly worse than the Multivariate CACE in terms of the power. But combined with its performance in controlling the type I error, the parametric bootstrap test is preferred for moderate sample sizes.

Therefore, the Multivariate CACE is the clear winner in terms of statistical power.

Estimating the type I error is considered as a special case in our power analysis. When the true value of the effect size equals 0, the type I error can be estimated via (4.11). Table 4.5 shows the results of $\hat{\alpha}$ for different variance-covariance matrices, sample sizes and CACE models. Ideally, the estimates of type I error should be approximately 0.05 as we assumed $\alpha = 0.05$. However, it is worth noticing that the estimates of α inflate for the multivariate model under moderate sample sizes. The occurrence of the mixture part in our Multivariate CACE model could be a possible explanation of this phenomenon. It could also be a result of a low compliance rate. We set p_c to 0.4 in our simulation so that only 40% of the data could be used to estimate the treatment effect. If the sample size equals 100, then the effective sample size is only 40. We also notice that the the magnitude of variance-covariance matrices is not related to the magnitude of $\hat{\alpha}$, thus we only consider the small variance-covariance matrices from now on. Surprisingly, the performance of multiple univariate models is quite good that the $\hat{\alpha}$ does not inflate. This might be caused by

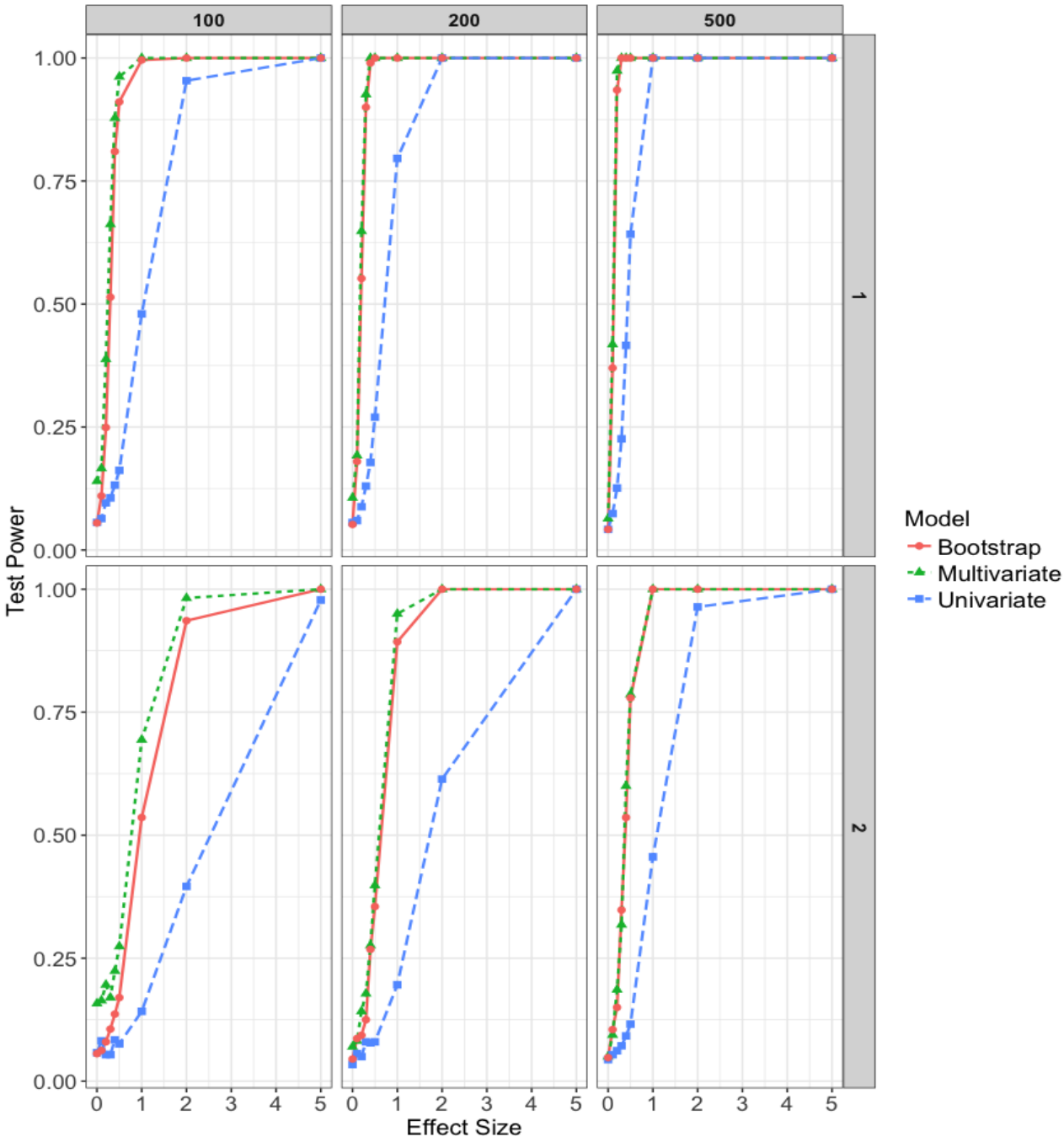

Figure 4.3: Power analysis based on 500 simulated datasets: the first row represents the results for smaller variance-covariance matrices ($\Sigma_{c_1}, \Sigma_{n_1}$) and the second row represents the results for larger variance-covariance matrices ($\Sigma_{c_2}, \Sigma_{n_2}$); the green dash curve represents the results for the Multivariate CACE, the red solid curve represents the results for the parametric bootstrap test and the blue dash curve represents the results for the Univariate CACE ($\delta_c = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5)$).

the Bonferroni correction we applied to multiple univariate tests. The Bonferroni correction tends to be a bit too conservative as it is trying to make it unlikely that you would make even one false rejection.

Theoretically, the LR test statistic G asymptotically follows the chi-square distribution under large sample sizes. Unfortunately, as we mentioned earlier, the approximation is inaccurate even for moderate sample sizes (N = 100, 200) for the multivariate model when the compliance rate is low. So we tried the parametric bootstrap test to address this issue and the results are presented in Table 4.5. To be consistent, we also performed the parametric bootstrap test for large sample size scenarios, which is not suggested in practice due to high computational cost. We could observe a clear improvement from Table 4.5 that $\hat{\alpha}$ drops to the nominal level even for the smallest sample size. This approach also performs well for large sample sizes.

Table 4.5: Estimates of type I error from the multivariate likelihood ratio test, the univariate likelihood ratio test, the parametric bootstrap test for different variance-covariance matrices.

Variance-covariance Matrix	1			2		
Model	Mul		Uni	Mul		Uni
N / Test	LR	Bootstrap	LR	LR	Bootstrap	LR
100	0.140	0.055	0.056	0.158	0.056	0.058
200	0.106	0.052	0.056	0.070	0.045	0.034
500	0.064	0.042	0.042	0.050	0.048	0.044
1000	0.048	0.053	0.032	0.048	0.052	0.038

We gathered the test statistics from 500 simulated datasets and visualized the results in Figure 4.4. Each panel shows the distribution of test statistics grouped by the sample size. For moderate sample sizes, the kernel estimated distribution has a lower peak and a fatter tail, and as sample size increases, the discrepancy disappears accordingly. As expected, a perfect overlap of two densities appears when sample size increases to 1000.

Another parameter of interest is the compliance rate. In order to figure out whether higher compliance rate leads to lower type I error, we chose several different values of p_c in our simulation study. Table 4.6 shows the estimates of α for $p_c = (0.4, 0.5, 0.6, 0.8)$. Under fixed sample size, as compliance rate increases from 0.4 to 0.6, $\hat{\alpha}$ decreases quickly. The decreasing slows down after the compliance rate reaches 0.6. To be consistent, we performed the parametric bootstrap test for different values of p_c as well. Upon closer inspection we can see that the parametric bootstrap test gives quite robust results that $\hat{\alpha}$ fluctuates within a small range for different sample sizes and compliance rates. These results have been visualized in the left panel in Figure 4.5.

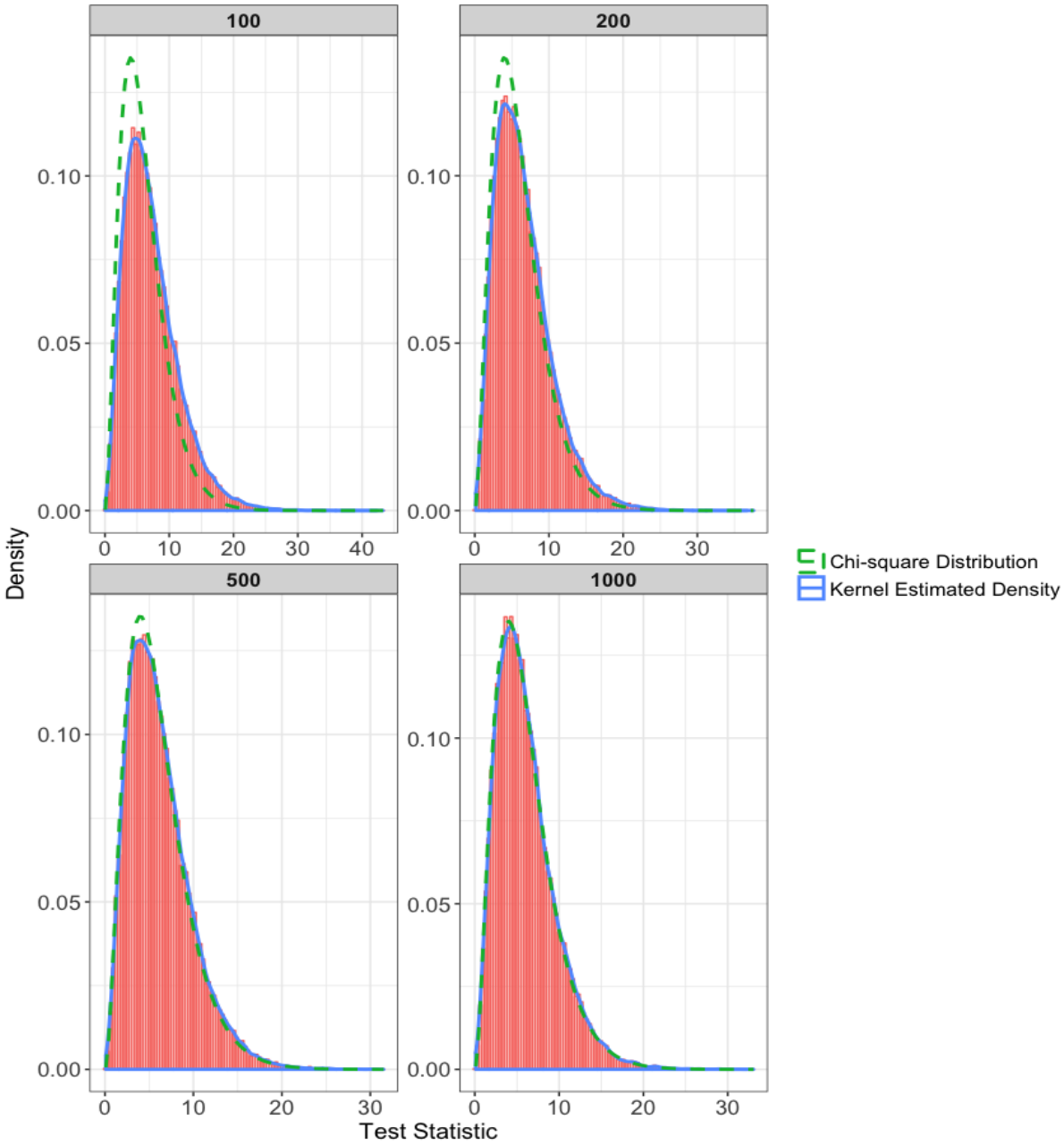

Figure 4.4: Distribution of LR test statistics: the green dash curve is the density function of the chi-square distribution and the blue solid curve is the kernel estimated density function from test statistics.

Table 4.6: Estimates of type I error for different compliance rates.

p_c	$p_c = 0.4$		$p_c = 0.5$		$p_c = 0.6$		$p_c = 0.8$	
N / Test	LR	Parametric Bootstrap	LR	Parametric Bootstrap	LR	Parametric Bootstrap	LR	Parametric Bootstrap
100	0.140	0.055	0.092	0.056	0.088	0.060	0.086	0.052
200	0.106	0.052	0.062	0.045	0.056	0.050	0.058	0.047
500	0.064	0.042	0.054	0.051	0.048	0.046	0.048	0.051

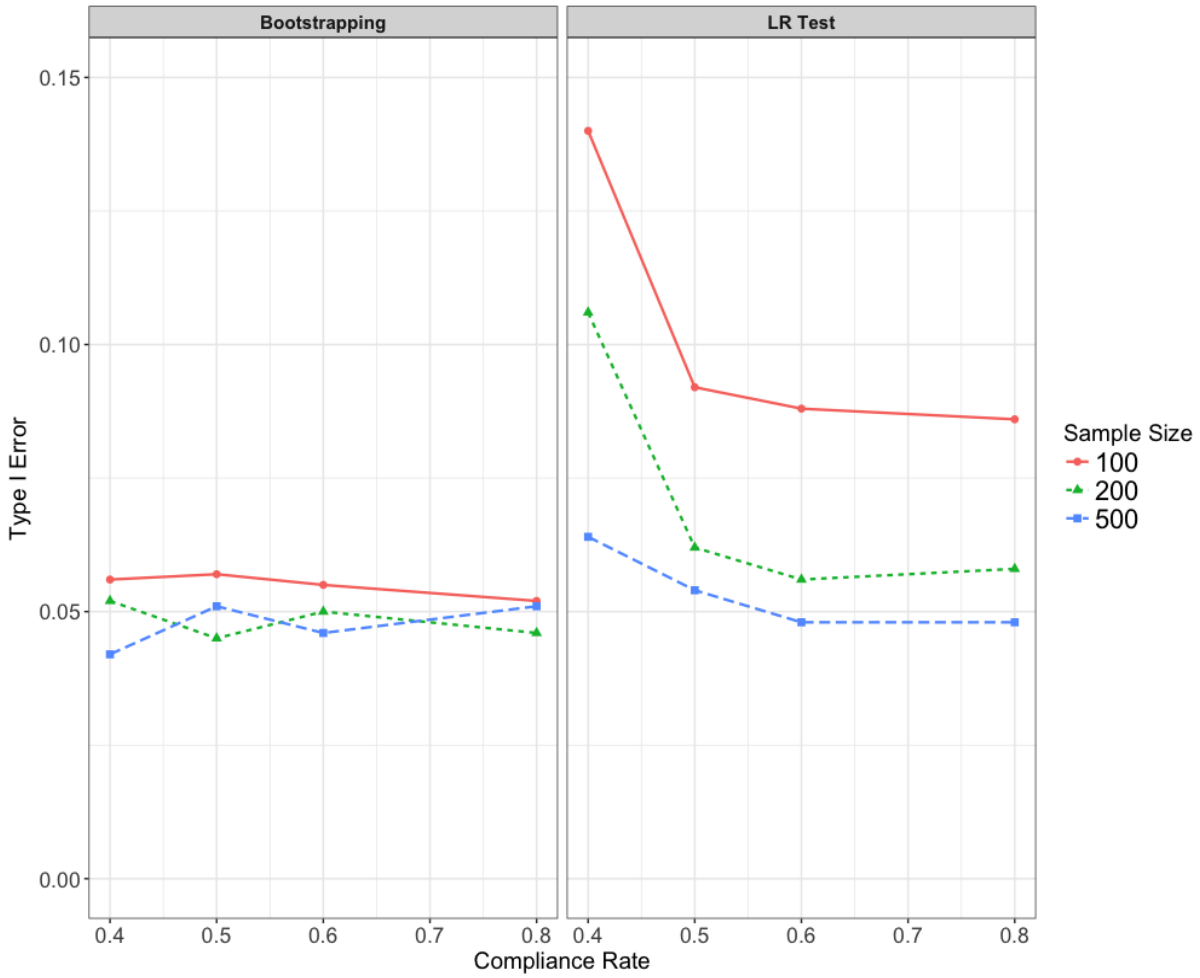

Figure 4.5: Estimated type I error for different compliance rates.

4.2.4 Simulation for Independent Outcomes

As mentioned in our motivation, independent outcomes can be treated as a special case in multivariate analysis. We would like to have a brief discussion on the results for independent multivariate outcomes in this section. Instead of using (4.1) and (4.2), we chose two diagonal matrices as variance-covariance matrices to generate datasets. The diagonal elements of two new variance-covariance matrices are the same as (4.1) and (4.2).

The distribution of estimates is shown in Figure 4.6, grouped by the sample size. Not surprisingly, two CACE models have comparable performance.

We also present the distribution of the length of confidence intervals in Figure 4.7. The width of confidence intervals decreases as sample size increases. We are happy to observe that the Multivariate CACE model still outperforms multiple Univariate models in the length of confidence intervals.

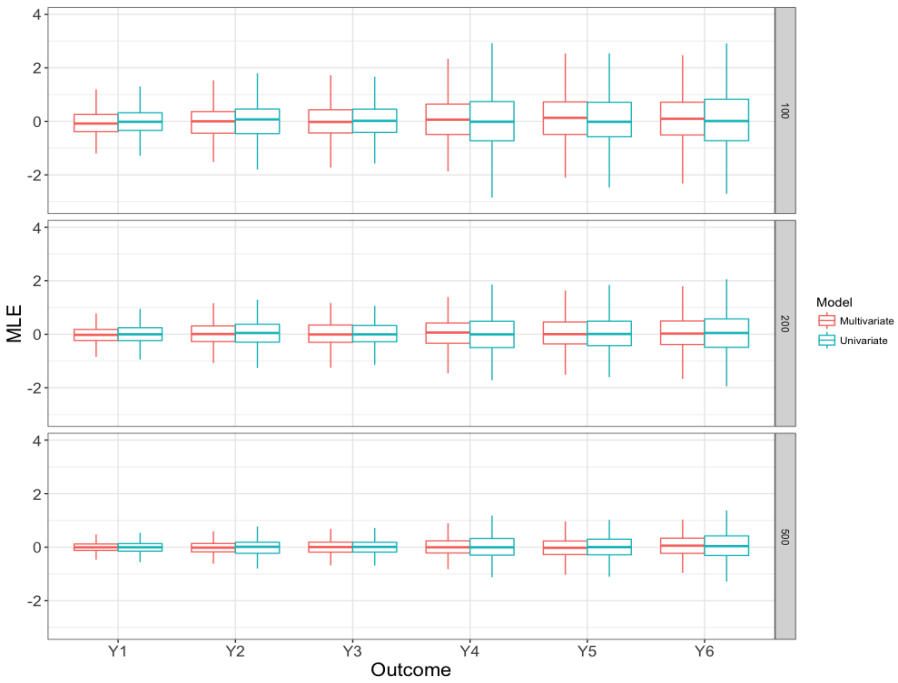

Figure 4.6: Distribution of MLEs for independent outcomes: red boxes represent the results for the Multivariate CACE and green boxes represent the results for the multiple Univariate CACE ($\delta_c = 0$).

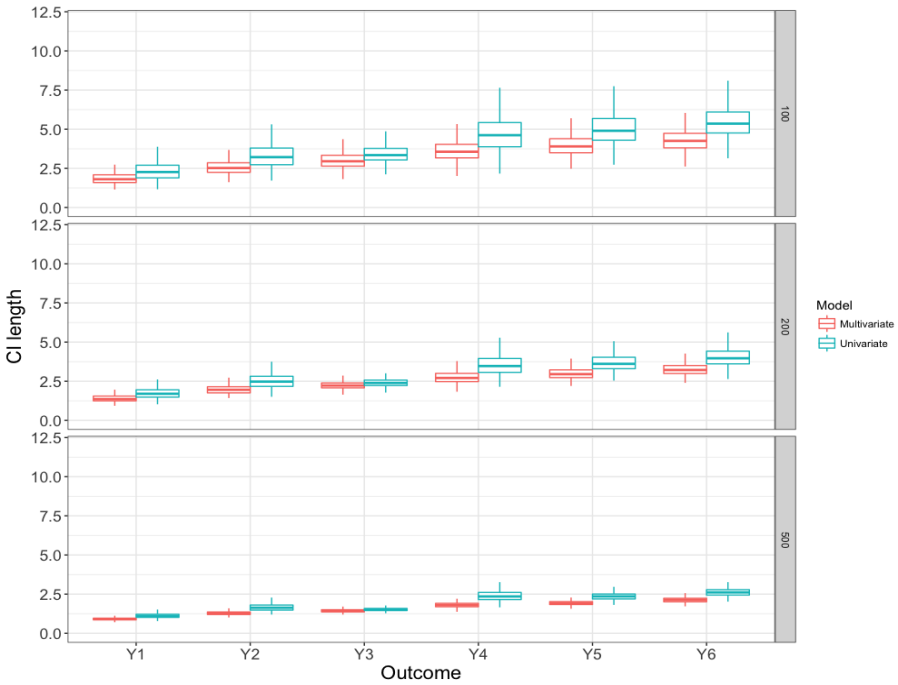

Figure 4.7: Distribution of the length of confidence intervals for independent outcomes: red boxes represent the results for the Multivariate CACE and green boxes represent the results for the multiple Univariate CACE.

Chapter 5

Application

We return to our motivating example described in Chapter 2. For consistency and completeness, we fitted both the Multivariate CACE and multiple Univariate CACE models to the Arthritis Health Journal data to compare the estimates of parameters. Since the estimated value of the compliance rate is around 0.4, and the sample size is only 80 in total, we decided to conduct both the global likelihood ratio test and the parametric bootstrap test to evaluate the significance of the treatment effect. Additionally, we also performed Intention-to-treat (ITT) analysis and As-treated (AT) analysis to our dataset. The ITT analysis is considered to be comparably conservative in estimating method effectiveness while the AT analysis may exaggerate the method effectiveness, so we expect to see the estimates obtained from the Multivariate CACE fall in the range formed by the estimates from the ITT analysis and the AT analysis.

5.1 Point Estimate

The estimates of δ_c are listed in Table 5.1. Upon closer inspection we can see the discrepancy between Multivariate CACE estimators and Univariate CACE estimators, which might be caused by the non-ignorable correlations among different health outcomes. Recall that all health outcomes are change scores from baseline and $\hat{\delta}_c$ is the estimate of the difference in change scores between treatment and control groups with the positive values indicating beneficial treatment effects of using the online tool except for outcome Y_5 whose positive change score suggests harmful treatment effects. It is worth noticing that the Multivariate CACE estimates for the 6 outcomes are all in the direction of beneficial treatment effects of the online tool whereas the Univariate CACE estimates for Y_2 and Y_3 point to the direction of harmful treatment effects of the online tool. The shrinkage of the extreme estimates in the Multivariate CACE helps to provide more precise estimates by borrowing information from multiply correlated data and is demonstrated to have impact on CACE estimates with potential clinical implications in this application.

Table 5.2 and Table 5.3 show the estimates of means for compliers in the treatment group and never-takers.

Table 5.1: Estimates of δ_c for the Multivariate CACE analysis, the Univariate CACE analysis, the ITT analysis and the AT analysis.

Outcome / Model	CACE		ITT	AT
	Mul	Uni		
Y_1	0.961	1.250	1.811	1.163
Y_2	2.017	-1.030	1.247	5.328
Y_3	4.008	-2.165	-0.499	3.883
Y_4	9.914	12.545	5.315	11.678
Y_5	-5.814	-11.030	-2.757	-10.140
Y_6	0.976	2.017	0.592	1.973

Table 5.2: Estimates of μ_c for the Multivariate CACE analysis, the Univariate CACE analysis, the ITT analysis and the AT analysis.

Outcome / Model	CACE		ITT	AT
	Mul	Uni		
Y_1	2.078	1.789	2.898	2.685
Y_2	0.383	3.430	0.573	3.652
Y_3	0.792	6.965	1.294	4.348
Y_4	0.753	-1.879	3.431	8.670
Y_5	-4.769	0.447	-3.382	-9.022
Y_6	1.315	0.275	0.708	1.773

Table 5.3: Estimates of μ_n for the Multivariate CACE analysis, the Univariate CACE analysis, the ITT analysis and the AT analysis.

Outcome / Model	CACE		ITT	AT
	Mul	Uni		
Y_1	1.428	1.486	1.087	1.522
Y_2	-1.080	-2.537	-0.674	-1.675
Y_3	0.853	-1.944	1.793	0.465
Y_4	-2.864	-2.051	-1.884	-2.982
Y_5	1.801	0.130	-0.625	1.118
Y_6	-0.514	-0.193	0.116	-0.200

Estimates of variances for the Univariate CACE are listed in (5.1) and (5.2). As shown in (5.3) and (5.4), the variance-covariance matrices for compliers and never-takers are quite similar. The variances of Y_1 to Y_5 are huge and the correlations among outcomes are non-negligible.

$$\hat{\sigma}_c = (70.478,\ 69.283,\ 166.042,\ 518.459,\ 273.333,\ 7.327), \tag{5.1}$$

$$\hat{\sigma}_n = (127.657,\ 228.597,\ 130.011,\ 78.076,\ 97.585,\ 2.312); \tag{5.2}$$

$$\hat{\Sigma}_c = \begin{bmatrix} 60.143 & 23.323 & 47.894 & 106.599 & -47.663 & 6.441 \\ 23.323 & 80.742 & 92.118 & 61.113 & -51.681 & 4.575 \\ 47.894 & 92.118 & 178.645 & 136.651 & -117.646 & 15.241 \\ 106.599 & 61.113 & 136.651 & 525.573 & -102.971 & 32.906 \\ -47.663 & -51.681 & -117.646 & -102.971 & 251.736 & -10.609 \\ 6.441 & 4.575 & 15.241 & 32.906 & -10.609 & 5.616 \end{bmatrix}, \tag{5.3}$$

$$\hat{\Sigma}_n = \begin{bmatrix} 127.752 & 74.798 & 53.929 & 65.222 & -75.603 & 9.506 \\ 74.798 & 217.553 & 123.435 & 42.303 & -90.385 & 11.536 \\ 53.929 & 123.435 & 148.105 & 38.099 & -65.327 & 4.778 \\ 65.222 & 42.303 & 38.099 & 103.726 & -58.588 & 11.084 \\ -75.603 & -90.385 & -65.327 & -58.588 & 119.076 & -9.910 \\ 9.506 & 11.536 & 4.778 & 11.084 & -9.910 & 2.690 \end{bmatrix}. \tag{5.4}$$

Another important parameter in our model is the compliance rate. As we discussed in Chapter 1, one limitation of the multiple Univariate CACE is that it yields 6 different estimates of p_c (see Table 5.4). We calculated the average value of these 6 estimates for

Table 5.4: Maximum likelihood estimates of p_c.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Mean
$\hat{p}_c$	0.45	0.43	0.44	0.43	0.45	0.39	0.43

interpretation purpose. The estimate of p_c provided by the Multivariate CACE is 0.38.

5.2 Interval Estimate

We compared the confidence intervals of the treatment effect δ_c for two CACE models (see Table 5.5 and Figure 5.1). An important observation is that the Univariate CACE gives wider confidence intervals than the Multivariate CACE for most outcomes. This confirms the finding from our simulation study that the Multivariate CACE outperforms the Univariate CACE in the length of confidence intervals. It is also worth noticing that all confidence intervals contain 0, which implies that the treatment effect is not statistically significant. The length of confidence intervals also reflects the magnitude of variances of outcomes.

Table 5.5: 95% simultaneous confidence intervals of δ_c for the Multivariate CACE and the Univariate CACE.

Outcome / Model	Mul CACE	Uni CACE
Y_1	(-6.796, 8.718)	(-8.688, 11.188)
Y_2	(-7.338, 11.371)	(-12.340, 10.340)
Y_3	(-9.872, 17.887)	(-18.800, 14.470)
Y_4	(-12.349, 32.177)	(-8.590, 33.681)
Y_5	(-21.215, 9.587)	(-27.430, 5.370)
Y_6	(-1.337, 3.289)	(-0.702, 4.737)

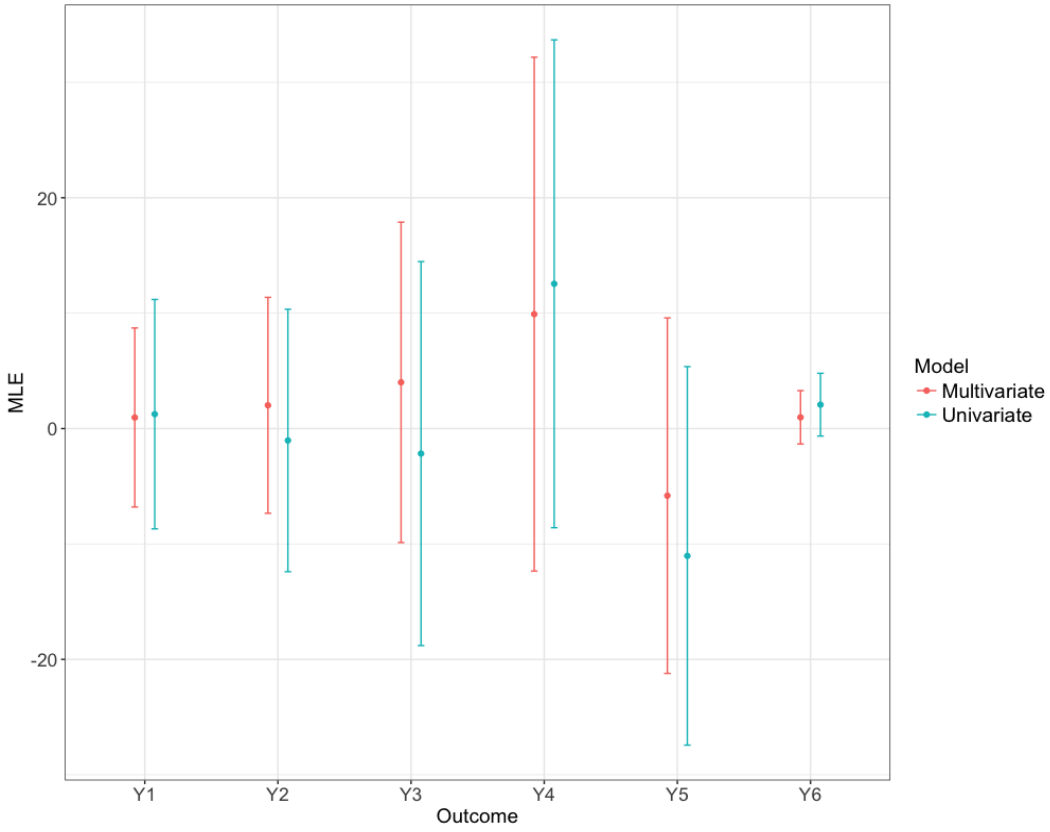

Figure 5.1: MLEs and the corresponding 95% confidence intervals: the red bars represent the Multivariate CACE and the green bars represent the Univariate CACE.

5.3 Hypothesis Test

The global likelihood ratio test was conducted for both CACE models, and the parametric bootstrap test was only conducted for the Multivariate CACE to control the inflation of type I error.

5.3.1 Multivariate CACE Analysis

In the Arthritis Health Journal data, we have 6 health outcomes in total. We consider the global likelihood ratio test first. Recall that the test statistic G is calculated as

$$G = -2\left(l_{reduced}|_{\hat{\theta}_r} - l_{full}|_{\hat{\theta}_f}\right), \tag{5.5}$$

where the reduced model sets δ_c to 0. Under the null hypothesis, G asymptoticly follows a chi-square distribution with degree of freedom 6.

Applying the calculation to the Arthritis Health Journal data, the test statistic equals 2.381 and the p value equals 0.881. Therefore, we fail to reject the null hypothesis; that is, the treatment effect is not statistically significant.

Following the steps described in Section 3.3.8, we conducted a parametric bootstrap test to get the estimated distribution for G (see Figure 5.2). It is clear that the kernel estimated density curve and the density of chi-square distribution overlap each other for most parts, but the differences still exist around the peak and the right tail. The kernel estimated density has a fatter tail and a lower peak as seen in the simulation study. We fail to reject the null hypothesis, which confirms the conclusion we have obtained from the global likelihood ratio test.

5.3.2 Univariate CACE Analysis

As for Univariate CACE analysis, Bonferroni correction should be considered for multiple tests. Therefore, the cut-off value for significance should be 0.0083 at level 0.05 when $k = 6$. Six tests were conducted separately and we obtained 6 p values as follows

$$0.739,\ 0.813,\ 0.734,\ 0.127,\ 0.076,\ 0.065.$$

Consequently, we fail to reject any of these six hypotheses.

Despite the discrepancy between the estimates, the Multivariate CACE model and multiple Univariate CACE models reach the same conclusion that the treatment effect is not statistically significant in the Arthritis Health Journal Study.

5.4 Effects of Baseline Covariates on Compliance Mechanism

There are 5 binary pre-treatment covariates in our Arthritis Health Journal dataset: Disease Time, Disease Activity I, Disease Activity II, Age and Gender. Disease Activity I and Disease Activity II are two different ways to define disease activity, so we just keep Disease Activity I as one of the predictors. As defined in (3.13), p_c is no longer a constant when

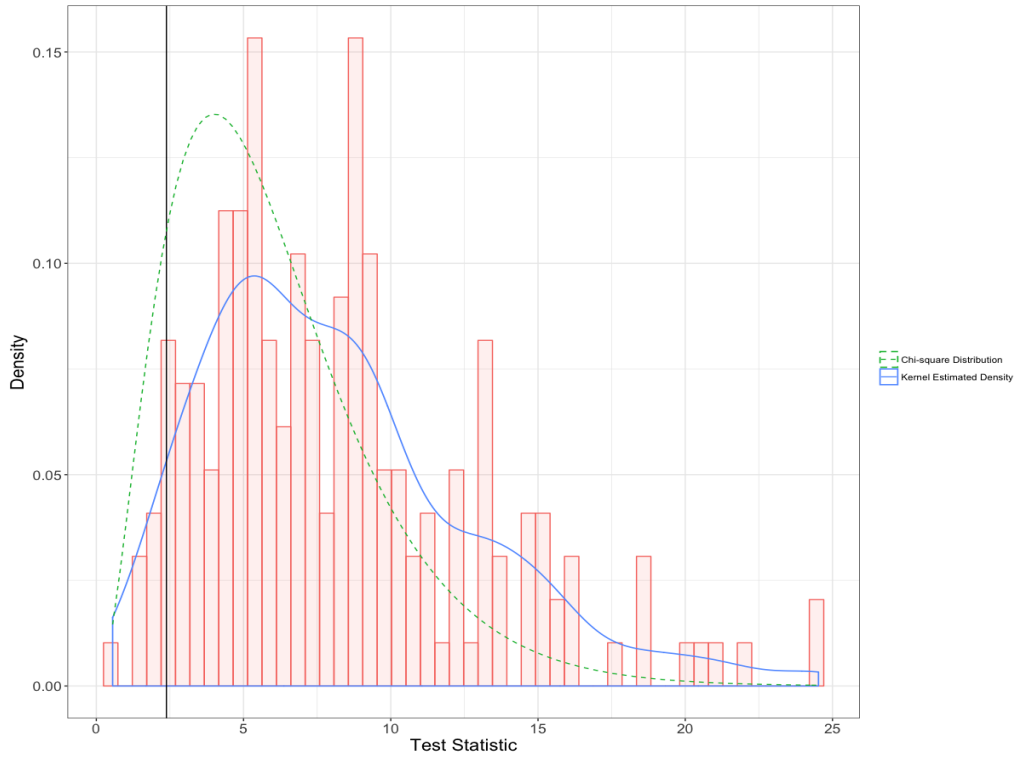

Figure 5.2: Estimated distribution of the test statistic via the parametric bootstrap test: the red bars represent counts, the blue solid curve is the kernel estimated density, the green dash curve is the density function of chi-square distribution with degree of freedom 6 and the black vertical line represents the value of the original LR test statistic G_0.

considering potential effects of baseline covariates on the compliance mechanism:

$$p_c = \Psi(c, w, \psi) = \frac{\exp(\psi_0 + \psi_1 w_1 + \psi_2 w_2 + \psi_3 w_3 + \psi_4 w_4)}{1 + \exp(\psi_0 + \psi_1 w_1 + \psi_2 w_2 + \psi_3 w_3 + \psi_4 w_4)}. \tag{5.6}$$

Table 5.6 shows how the estimates were affected by adding baseline covariates to our CACE models. The effect of baseline covariates on the estimates is somewhat limited that even the estimate of p_c did not change much after adding baseline covariates to our models.

In addition to the new estimates of parameters, we also obtained a new LR test statistic G and the corresponding p value: $G = 2.749$ and p value = 0.840. Considering baseline covariates does not change our conclusion that the treatment effect is not statistically significant.

Table 5.7 shows the estimated intercept and coefficients to predict the compliance behavior. Based on the calculated p value, only the Disease Activity I and Age are significant predictors at level 0.05. Recall that w_1 is the binary indicator for Disease Duration; w_2 is binary indicator for Disease Activity I; w_3 is the binary indicator for Gender; and w_4 is the binary indicator for Age, we consider the interpretation of the intercept and coefficients.

Table 5.6: Comparison of new and old estimates of δ_c & p_c.

Outcome / Model	Multivariate CACE		Univariate CACE	
	Old	New	Old	New
Y_1	0.961	0.301	1.250	-0.343
Y_2	2.017	2.436	-1.030	-2.668
Y_3	4.008	4.952	-2.165	-2.634
Y_4	9.914	9.345	12.545	12.327
Y_5	-5.814	-6.165	-11.030	-11.741
Y_6	0.976	0.953	2.017	1.876
p_c	0.380	0.386	0.430	0.401

Table 5.7: Estimates of the intercept and coefficients.

	MLE	se	z value	p value
Intercept (ψ_0)	-1.290	0.785	-1.644	0.100
Disease Duration (ψ_1)	-0.768	1.043	-0.736	0.461
Disease Activity I (ψ_2)	1.658	0.718	2.310	0.021
Gender (ψ_3)	-0.761	1.543	-0.493	0.622
Age (ψ_4)	-0.842	0.358	-2.349	0.019

The intercept ψ_0 is the log odds of being a complier for a female patient with late disease, low disease activity under 54.5 years old. Therefore, $\hat{\psi}_0 = -1.290$ suggests that a young woman with late disease and mild symptoms is less likely to be a complier. The coefficients could be interpreted as the difference in log odds ratio with other covariates fixed. For example, ψ_1 is the difference in log odds of being a complier between patients with early disease and patients with late disease when the disease activity, gender and age are fixed. In conclusion, young female patients with late disease and high disease activity 1 are more likely to be compliers.

Though we fail to observe huge difference between with and without the baseline covariates, adding baseline covariates to our model still has some scientific meaning for researchers. For instance, researchers may be capable of predicting participants' compliance behaviors based on their baseline measurements and take some actions to avoid noncompliance behavior.

Chapter 6

Discussion

6.1 Summary

Many researchers have encountered the noncompliance issue in the RCTs involving human subjects when they try to evaluate the efficacy of a new treatment. The Intention-to-treat (ITT) analysis only concentrates on the initial treatment assignment, while the as-treated (AT) analysis concerns the actual receipt of treatment and ignores the assignment. With the occurrence of non-compliers, both the ITT analysis and the AT analysis would generate biased estimates of method effectiveness. CACE analysis was introduced to provide an unbiased estimate of the method effectiveness. Due to the complexity of treatment effect evaluation, correlated multivariate outcomes are common in epidemiological studies. In this thesis, we generalized the Univariate CACE model to the Multivariate CACE model in order to better deal with multivariate health outcomes. Instead of analyzing multivariate outcomes separately, we managed to model the multiple outcomes all together without losing key information.

We conducted a systematic simulation study to compare Multivariate and Univariate CACE models in different scenarios. We also applied both models to the Arthritis Health Journal data to estimate the treatment effect. In particular, we performed a global likelihood ratio test to test the significance of the treatment effect as well as a parametric bootstrap test to control type I error inflation. Interestingly, we found the Multivariate CACE estimates all pointed to the direction of beneficial effects of using the online tool whereas two out of six estimates in the Univariate CACE pointed to the direction of harmful treatment effects. Despite this difference in estimates, the treatment effect is not statistically significant in the Arthritis Health Journal Study according to the results of both tests.

6.1.1 Interesting Findings from Simulation Study

We summarize some interesting findings from our simulation study, which shows the clearer evidence of advantages of the Multivariate CACE model over multiple Univariate CACE models.

First of all, one of the advantages of the Multivariate CACE model over multiple Univariate CACE models is that it produces more precise estimates with smaller variance of the MLEs, narrower confidence intervals and higher coverage rate. The multivariate model further expands its advantage as the variance of the outcomes increases. Even in cases where the multivariate outcomes are independent from each other, the Multivariate CACE still performs slightly better than the univariate model.

Secondly, performing multivariate analysis could bring considerable improvement in statistical power. As shown in Figure 4.3, the test power increases by 0.7 at most by using the multivariate model when sample size equals 100 with smaller variance-covariance matrices. Given the same effect size and variance-covariance matrices, the difference in power between two models becomes larger as sample size increases. Additionally, it seems very promising that the power of the Multivariate CACE model grows more faster and reaches 1 earlier as effect size and sample size increase.

However, we lost the control of the type I error that should be 0.05 in theory when modeling multivariate outcomes via the Multivariate CACE model. Parametric bootstrap test was conducted to address this issue. It turns out that the parametric bootstrap test is very robust to control both the type I error and the type II error simultaneously for different values of the sample size and compliance rate. Figure 4.4 shows the estimated distribution of LR test statistic from which we could observe the perfect overlap of the estimated distribution and the chi-square distribution when sample size is large enough.

6.1.2 Limitations

Despite the fact that the Multivariate CACE outperforms the Univariate CACE in estimation and statistical power, several important issues are not addressed.

We assumed the health outcomes to follow a multivariate normal distribution in our simulation study and in the Arthritis Health Journal Study. In fact, the multivariate health outcome does not always follow an exact multivariate normal distribution, which may violate the assumptions of the likelihood function. A nonparametric model can be a better candidate to model the non-normal outcomes, however it is beyond the scope of this project.

In the Arthritis Health Journal Study, participants who have used the online tool for no less than 3 times are defined as compliers. As compared to the never-takers in the control group, never-takers in the treatment group may use the online tool for once or twice. The definition of compliers implies no difference between never-takers in two groups. Thus, the exclusion restriction assumption is violated. An alternative definition of the compliers in the AHJ Study may make the assumption of exclusion restriction more reasonable, which will be investigated in the future.

In most clinical trials, a moderate sample size of 100 or 200 can be very common. To control the inflation of type I error, the parametric bootstrap test is required. The paramet-

ric bootstrap test performs well in controlling both the type I error and the type II error at the sacrifice of significantly increased computational cost. Therefore, it may not be attractive to spend hours on a simple test. But we have not found a more efficient way to control both error rates simultaneously.

6.2 Future Work

The Multivariate CACE model proposed in this thesis is a rather basic one, as it ignores the effects of the baseline covariates on health outcomes. We only considered the effect of baseline covariates on the compliance rate for the Arthritis Health Journal data. Further improvements could be made to the Multivariate CACE model to provide more accurate analysis results by considering the baseline characteristics.

Our proposed Multivariate CACE model highly relies on some crucial assumptions, further work can be done to explore the scenarios when one or few assumptions are violated. The random assignment is not a realistic assumption when involving human subjects due to ethical concerns. Furthermore, if the SUTVA and the exclusion restriction assumption are violated, the CACE model would not be identifiable. In future, we will propose a more robust multivariate CACE model or other nonparametric models to address above concerns.

Bibliography

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.

Arin M. Connell. Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. *The American Journal of Drug and Alcohol Abuse*, 35(4):253–259, 2009. URL https://doi.org/10.1080/00952990903005882.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

Richard F. Haase and Michael V. Ellis. Multivariate analysis of variance. *Journal of Counseling Psychology*, 34(4):404 – 413, 1987. ISSN 0022-0167.

Keisuke Hirano, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000. doi: 10.1093/biostatistics/1.1.69. URL http://dx.doi.org/10.1093/biostatistics/1.1.69.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Guido W. Imbens and Donald B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997. ISSN 00905364. URL http://www.jstor.org/stable/2242722.

Christopher R. Knox, Ranjit Lall, Zara Hansen, and Sarah E. Lamb. Treatment compliance and effectiveness of a cognitive behavioural intervention for low back pain: a complier average causal effect approach to the best data set. *BMC Musculoskeletal Disorders*, 15(1):17–27, 2014. URL https://doi.org/10.1186/1471-2474-15-17.

Young Jack Lee, Jonas H. Ellenberg, Deborah G. Hirtz, and Karin B. Nelson. Analysis of clinical trials by treatment actually received: Is it really an option? *Statistics in Medicine*, 10(10):1595–1605, 1991. ISSN 1097-0258. doi: 10.1002/sim.4780101011. URL http://dx.doi.org/10.1002/sim.4780101011.

MD LEWIS b. Sheiner and Donald B. Rubin. intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics*, 57(1):6–15, 1995.

Mary J. Lindstrom and Douglas M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988. ISSN 01621459. URL http://www.jstor.org/stable/2290128.

CJ McDonald, SL Hui, and WM Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *M.D. computing : computers in medical practice*, 9(5):304–312, 1992. ISSN 0724-6811. URL http://europepmc.org/abstract/MED/1522792.

JosÃl' C. Pinheiro and Douglas M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996. ISSN 1573-1375. URL https://doi.org/10.1007/BF00140873.

Catherine Stanger, Stacy R. Ryan, Hongyun Fu, and Alan J. Budney. Parent training plus contingency management for substance abusing families: A complier average causal effects (cace) analysis. *Drug and Alcohol Dependence*, 118(2):119 – 126, 2011. URL http://www.sciencedirect.com/science/article/pii/S0376871611001335.