CLARIFICATION AND IDENTIFICATION OF CAUSAL ESTIMANDS USING PRINCIPAL STRATUM STRATEGY IN CLINICAL TRIALS WITH TWO ACTIVE TREATMENTS

By

Hanyue Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Epidemiology - Doctor of Philosophy

2021

ABSTRACT

CLARIFICATION AND IDENTIFICATION OF CAUSAL ESTIMANDS USING PRINCIPAL STRATUM STRATEGY IN CLINICAL TRIALS WITH TWO ACTIVE TREATMENTS

By

Hanyue Li

The randomization process in clinical trials is disrupted with the existence of post-randomization events or intercurrent events. The "gold-standard" intention-to-treat (ITT) estimand therefore loses its clinical relevance because it compares treatment assignments instead of actual received treatments. Alternative estimands need to be defined and identified to quantify the causal effect of treatments. In this dissertation, we focus on the intercurrent event of treatment nonadherence and identify the causal estimands in clinical trials with two active treatments.

We work under the Neyman-Rubin causal framework and the principal stratification framework. First, we propose a nonparametric approach which identifies the complier average causal effect (CACE) as the ratio of the ITT effect of treatment assignment on the outcome to the ITT effect of treatment assignment on the treatment received under the exclusion restriction, monotonicity, and no partial-compliers assumptions. We discuss violation of the identification assumptions and derive the corresponding bias formulas. Simulations with various degrees of assumption violations are conducted to evaluate the performance and sensitivity of the approach. The results show that the nonparametric approach can yield an unbiased estimator for CACE when sample size is 500 or above and the percentage of compliers is above or equal to 70%. In addition, increasing the number of compliers has the potential to reduce the bias to as close as zero.

Second, we propose a multisite design approach which identifies the CACE under the zero correlation assumption. We derive the bias formula when measurement errors and omitted variables exist. Simulations across various scenarios are conducted for the oracle, naïve, and

bootstrap estimators. The results show that multisite design approach can yield an unbiased estimator if there are no measurement errors and omitted variables. Increasing the number of people in each site can reduce the bias because it reduces the variation of the measurement errors. Increasing the number of sites, on the other hand, does not make a significant impact on the bias.

We apply the two proposed approaches to the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) data to identify the causal effects of the medication augmentation and the causal effects of the medication switching. We show that augmenting citalopram with sustained-release bupropion is no better than augmenting citalopram with buspiron in terms of remission but has a higher response rate. Augmenting citalopram with sustained-release bupropion also reduces the 17-item Hamilton Rating Scale for Depression ($HAM\text{-}D_{17}$) score greater than augmenting citalopram with buspiron. We also show that switching to extended-release venlafaxine has a slightly better performance compared switching to sustained-release bupropion in terms of remission, $HAM\text{-}D_{17}$ scores at the end of the study, and reduction of $HAM\text{-}D_{17}$ scores.

To Mom and Dad.
Thank you for everything.
I love you to the moon and back.

ACKNOWLEDGEMENTS

A journey is approaching to the destination.

It is a journey filled with adventures, challenges, confusions, and tears.

It is also a journey full of gains, reconciliations, appreciations, and happiness.

At the end of the journey, I would like to express my sincere gratitude to my advisor and dissertation committee chair, Dr. Zhehui Luo. Dr. Luo guides me throughout this dissertation as well as my Ph.D. life with her endless knowledge, thoughtful insights and unlimited patience. This dissertation will not be completed without her support and advice.

I would like to thank my other committee members: Dr. Ling Wang, Dr. Ahnalee Brincks, and Dr. Joseph Gardiner for their kindness, encouragement, and the invaluable suggestions and comments they give.

I would like to thank the staff and faculty in my department who create such a friendly environment for us.

I would like to thank all the friends I've met along this journey.

Lastly, thank you my husband, Shitong Gu, for always staying by my side.

Thank you my mom and dad, Yun Lou and Wengang Li, for loving me and supporting me unconditionally all the time.

Thank you Audrey Gu, for being my daughter. Mommy loves you, always has, always will.

Thank you, beautiful journey.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xii

CHAPTER 1 INTRODUCTION AND AIMS 1

CHAPTER 2 BACKGROUND 4
2.1 Neyman-Rubin Causal Model 4
2.2 Principal Stratification Framework 6
2.3 Identification of Causal Estimands in RCTs with Treatment Nonadherence 7

CHAPTER 3 A NONPARAMETRIC APPROACH 12
3.1 Study Setup and Causal Estimands 12
3.2 Identification Strategies of the Causal Estimand 14
3.3 Sensitivity of the Causal Estimand to Key Assumptions 18
3.3.1 Violation of Assumption I (Exclusion Restriction) 19
3.3.2 Violation of Assumption II (Monotonicity) 20
3.3.2.1 Violation of Assumption II.a (No Irrationalists) 21
3.3.2.2 Violation of Assumption II.b (No Flip-Floppers) 23
3.3.3 Violation of Assumption III (No Partial-Complier) 24
3.4 Simulation Studies 26
3.4.1 Simulation Setup 26
3.4.2 Performance Metrics 30
3.4.3 Results 31
3.4.3.1 Performance Analysis 31
3.4.3.2 Sensitivity Analysis 33

CHAPTER 4 A MULTISITE DESIGN APPROACH 39
4.1 Identification Strategies of the Causal Estimand 40
4.2 Estimation of the Causal Estimand 44
4.3 The Problems of Omitted Variables and Measurement Errors 46
4.4 Simulation Studies 49
4.4.1 Simulation Setup 49
4.4.2 Results 53
4.4.2.1 Performance Analysis 53
4.4.2.2 Sensitivity Analysis 59

CHAPTER 5 APPLICATION 66
5.1 Medication Augmentation: CIT+BUP vs. CIT+BUS 69
5.1.1 Assessment of Assumptions 72
5.1.2 Analysis 79
5.1.3 Results 82

5.2 Medication Switching: BUP vs. VEN 85
5.2.1 Assessment of Assumptions 88
5.2.2 Results 93
5.3 Discussion 96

CHAPTER 6 DISCUSSION AND FUTURE WORK 100

APPENDICES 103
APPENDIX A Results of the Binary Outcomes for the Nonparametric Approach 104
APPENDIX B Results of the Binary Outcomes for the Multisite Design Approach 109

BIBLIOGRAPHY 115

LIST OF TABLES

Table 2.1 Principal strata classified by A(0) and A(1) for placebo-controlled RCTs. 7

Table 3.1 Principal strata classified by A(1) and A(2) for RCTs with two active treatments…....12

Table 3.2 Principal strata classified by A(1) and A(2) for RCTs with two active treatments under Assumption II, the monotonicity assumption, and Assumption III, the no partial-compliers assumption. 17

Table 3.3 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption II.a, the no irrationalists assumption, is violated. 21

Table 3.4 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption II.b, the no flip-floppers assumption, is violated. 23

Table 3.5 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption III, the no partial-compliers assumption, is violated. 25

Table 3.6 Summary of the parameter values used to generate the potential outcomes for the nonparametric approach. Scenario A is when all assumptions are satisfied. Scenario B is when Assumption I, the exclusion restriction assumption, is violated. Scenario C is when Assumption II.a, the no irrationalists assumption, is violated. Scenario D is when Assumption II.b, the no flip-floppers assumption, is violated. Scenario E is when Assumption III, the no partial-compliers assumption, is violated. 29

Table 4.1 Principal strata classified by A(1) and A(2) for RCTs with two active treatments under Assumption II, the monotonicity assumption, and Assumption IV, the no never-takers assumption. 41

Table 4.2 Summary of the parameter values used to generate the potential outcomes for the multisite design approach. 52

Table 4.3 Sensitivity of the multisite design estimator across the nine scenarios based on number of sites and site sizes for the continuous outcome when site size is fixed if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$. 60

Table 4.4 Sensitivity of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites for the continuous outcome when site size is varied if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$..... 62

Table 5.1 Baseline and level 2 entry characteristics of the patients in the medication augmentation strategy. 70

Table 5.2 Observed treatment adherence in the medication augmentation strategy..... 72

Table 5.3 Side effects in level 2 in the medication augmentation strategy..... 74

Table 5.4 Population structure when Assumptions I (exclusion restriction), Assumption II (monotonicity), and Assumption III (no partial-compliers) are satisfied for the nonparametric approach. 75

Table 5.5 Population structure when Assumptions I (exclusion restriction), Assumption II (monotonicity), and IV (no never-takers) are satisfied for the multisite design approach. 76

Table 5.6 Reasons for missing HAM-D_{17} scores in the medication augmentation strategy..... 80

Table 5.7 Baseline and level 2 entry characteristics of the patients in the medication switching strategy. 86

Table 5.8 Observed treatment adherence in the medication switching strategy. 88

Table 5.9 Side effects in level 2 in the medication switching strategy. 90

Table 5.10 Reasons for missing HAM-D_{17} scores in the medication switching strategy. 93

Table 5.11 Population structure when Assumption III (no partial-compliers) is violated for the nonparametric approach in the medication augmentation strategy: situation 1. 97

Table 5.12 Population structure when Assumption III (no partial-compliers) is violated for the nonparametric approach in the medication augmentation strategy: situation 2. 97

Table B.1 Sensitivity of the multisite design estimator across the nine scenarios based on number of sites and site sizes for the binary outcome when site size is fixed if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$. 111

Table B.2 Sensitivity of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites for the binary outcome when site size is varied if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$..... 113

LIST OF FIGURES

Figure 2.1 Causal diagram for RCTs with binary actions and binary compliance status. Z is the randomly assigned treatment. A is the actually taken treatment. Y is the observed outcome. U represents the confounding variables that are either observed or unobserved and affect both A and Y. 4

Figure 3.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the nonparametric estimator across proportions of compliers for the continuous outcome when all assumptions are satisfied (scenario A). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.. 32

Figure 3.2 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of compliers for the continuous outcome when Assumption I, the exclusion restriction assumption, is violated (scenario B). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 35

Figure 3.3 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of irrationalists for the continuous outcome when Assumption II.a, the no irrationalists assumption, is violated (scenario C). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 36

Figure 3.4 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of flip-floppers for the continuous outcome when Assumption II.b, the no flip-floppers assumption, is violated (scenario D). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 37

Figure 3.5 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of partial-compliers for the continuous outcome when Assumption III, the no partial-compliers assumption, is violated (scenario E). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 38

Figure 4.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the nine scenarios based on number of sites s and site sizes n for the continuous outcome when site size is fixed. For example, label "s=50, n=25" indicates that the scenario has 50 sites and 25 individuals in each site. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal

strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities. 55

Figure 4.2 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites *s* for the continuous outcome when site size is varied. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities. 58

Figure 5.1 Linear regressions of site-specific estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on the site-level aggregated covariates in the medication augmentation strategy. 78

Figure 5.2 Estimated complier average causal effects (CACEs) and their 95% confidence intervals for the primary endpoint and secondary endpoints in the medication augmentation strategy. In each panel, the horizontal solid line represents the null. The dashed vertical line separates the nonparametric estimates on the left-hand side from the multisite design estimates on the right-hand side. The blue square represents main analysis. The green circle represents sensitivity analysis 1. The red triangle represents sensitivity analysis 2. The orange diamond represents sensitivity analysis 3. 83

Figure 5.3 Linear regressions of site-specific estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on the site-level aggregated covariates in the medication switching strategy. 92

Figure 5.4 Estimated complier average causal effects (CACEs) and their 95% confidence intervals for the primary endpoint and secondary endpoints in the medication switching strategy. In each panel, the horizontal solid line represents the null. The dashed vertical line separates the nonparametric estimates on the left-hand side from the multisite design estimates on the right-hand side. The blue square represents main analysis. The green circle represents sensitivity analysis 1. The red triangle represents sensitivity analysis 2. The orange diamond represents sensitivity analysis 3. 95

Figure A.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the nonparametric estimator across proportions of compliers for the binary outcome when all assumptions are satisfied (scenario A). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 104

Figure A.2 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of compliers for the binary outcome when Assumption I, the exclusion restriction assumption, is

violated (scenario B). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 105

Figure A.3 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of irrationalists for the binary outcome when Assumption II.a, the no irrationalists assumption, is violated (scenario C). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 106

Figure A.4 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of flip-floppers for the binary outcome when Assumption II.b, the no flip-floppers assumption, is violated (scenario D). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 107

Figure A.5 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of partial-compliers for the continuous outcome when Assumption III, the no partial-compliers assumption, is violated (scenario E). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$. 108

Figure B.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the nine scenarios based on number of sites s and site sizes n for the binary outcome when site size is fixed. For example, label "s=50, n=25" indicates that the scenario has 50 sites and 25 individuals in each site. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities. 109

Figure B.2 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites s for the binary outcome when site size is varied. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities. 110

CHAPTER 1

INTRODUCTION AND AIMS

Different estimands correspond to different types of treatment effects and can address fundamentally different research questions. In practice, however, the choices of estimands are often implied through subsequent analyses, leading to a consequence that the estimands misalign with the study objectives (Rubin, 2005). The same issue arises in clinical trials with post-randomization events. Clinical trials, usually randomized controlled trials (RCTs), are the foundation to provide valid estimates of the causal effects because randomization balances the observed and unobserved baseline confounding between treatment groups. To estimate the causal effects of treatments, the top-choice estimand is always the intention-to-treat (ITT) estimand. Nevertheless, with the existence of the post-randomization events such as loss to follow up or treatment nonadherence, the gold-standard ITT estimand is challenged as it merely compares the treatment assignments instead of the actual treatments taken, which loses clinical relevance and conflicts with the research question.

In 2010, the National Research Council (NRC) published a report on "*Prevention and Treatment of Missing Data in Clinical Trials*", emphasizing the importance of clarifying estimands before trial design and analyses with a focus on missing data (National Research Council, 2010). Last year, beyond the discussion of missingness, the International Council for Harmonisation (ICH) finalized the draft addendum to its E9 guideline on "*Estimands and Sensitivity Analysis in Clinical Trials*", further focusing on structuring a framework to define the estimand and clarifying the treatment effects of interest (ICH E9 working group, 2020). It is essential to have a clear understanding and description of the estimands and the causal effects of interest. Alternative

estimands besides ITT estimand, aiming at quantifying the causal effects of treatments, need to be considered for causal inference when post-randomization events exist.

In this dissertation, we will explore and clarify the potential causal estimands that can be defined in clinical trials with two active treatments when there exists treatment nonadherence and develop the identification strategies to estimate the corresponding causal effects. Our study is motivated by the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study which aims to evaluate the comparative effectiveness of various treatment options for participants with nonpsychotic major depressive disorder (MDD) who fail to have a satisfactory response to the first-line medication citalopram (Rush et al., 2004). Rush et al., (2006) compared treatment switching options: sustained-release bupropion (bupropion-SR), sertraline, or extended-release venlafaxine (venlafaxine-XR) and found insignificant differences among these treatments with respect to remission or response rates. Trivedi et al. (2006) and Bech et al. (2012) showed a slightly better performance of bupropion-SR than buspirone when comparing treatment augmentation options. These studies focused on the ITT estimand among two or more active treatments. Yet questions pertaining to nonadherence with treatments and what causal effects can be identified for which population remain to be discussed.

Nonadherence has been well recognized in the placebo-controlled trials and identification of the causal effects has been widely discussed (Angrist et al., 1996; Balke & Pearl, 1997; Imbens & Angrist, 1994; Imbens & Manski, 2004; Jiang et al., 2016; Little et al., 2009; Miratrix et al., 2018; Yuan et al., 2018). The definition and identification becomes more complicated if the comparison occurs between two active treatments rather than between a treatment and a placebo. Roy et al. (2008) and Long et al. (2010) identified the CACE for trials with more than one active treatment. However, their identification was built on parametric models. Yuan et al. (2018) utilized

the nature of multisite design which automatically creates multiple instruments from the site-by-treatment interactions, yet their study was essentially a placebo-controlled trial. We extend the methods by Angrist et al. (1996) and Yuan et al. (2018) to our two-active-treatments setting and define CACE without assuming any parametric models. Specifically, we will point-identify our causal estimand of interest rather than partial-identification using bounds (Swanson et al., 2018). We will work under the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) and the principal stratification framework (Frangakis & Rubin, 2002). Our specific aims are:

Aim 1: Define and identify the causal estimand and develop the identification strategies for randomized controlled trials with two active treatments subject to treatment nonadherence.

Specific aim 1.1: Define the identification strategies to estimate the corresponding causal effect through a nonparametric approach.

Specific aim 1.2: Define the identification strategies to estimate the corresponding causal effect via a multisite design approach.

Aim 2: Evaluate the performance of the estimand and assess its sensitivity when there are deviations from key identification assumptions through simulation studies.

Specific aim 2.1: Evaluate the performance and sensitivity of the estimand in the nonparametric approach when the exclusion restriction assumption or the two structural assumptions are violated through simulation studies.

Specific aim 2.2: Evaluate the performance and sensitivity of the estimand in the multisite design approach when the zero-correlation assumption is violated through simulation studies.

Aim 3: Apply the two proposed approaches to the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial to identify the causal effects of two medication augmentation options and the causal effects of two medication switching options.

CHAPTER 2

BACKGROUND

2.1 Neyman-Rubin Causal Model

We work under the potential outcome framework, which is also known as the Neyman-Rubin Causal Model. For a unit, the causal effect of an action relative to the other is the difference between the outcome had the unit been assigned to one action and the outcome had the unit been assigned to the other action (Rubin, 1974). Each action of the unit corresponds to one potential outcome. However, only the potential outcome associated with the action taken is actually observed, whereas the other is not realized (Holland, 1986; Imbens & Rubin, 2015). Accordingly, we can never observe the causal effect of one action over the other at the unit level.

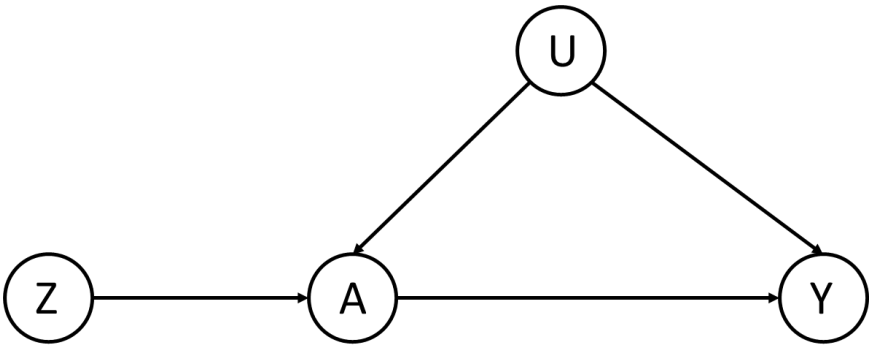

Figure 2.1 Causal diagram for RCTs with binary actions and binary compliance status. Z is the randomly assigned treatment. A is the actually taken treatment. Y is the observed outcome. U represents the confounding variables that are either observed or unobserved and affect both A and Y.

Formally, consider a placebo-controlled RCT with binary actions and binary compliance status whose causal diagram is shown in Figure 2.1. For unit i, let Z_i be the treatment randomly assigned, A_i be the treatment actually taken, Y_i be the observed outcome, and U_i be the confounding variables that are either observed or unobserved and affect both A_i and Y_i. We denote $Y_i(a)$ as the potential outcome if the unit takes treatment a where $a = 1$ for treatment and $a = 0$

for placebo. Similarly, we denote $A_i(z)$ as the potential treatment taken if the unit is assigned to treatment z, for $z \in \{0, 1\}$. Each unit has two potential outcomes $Y_i(1)$ and $Y_i(0)$, but only one of them can be observed. The individual causal effect of treatment is defined as $Y_i(1) - Y_i(0)$, which is generally impossible to obtain (Hernán & Robins, 2019). As we aim to identify the causal effects for a population of interest, we must rely on multiple units to estimate the average causal effects. We define the average causal effects as the difference in average potential outcomes between units taking one treatment and the same units had they taken the other for a certain population, e.g., $E\{Y(1) - Y(0)\}$.

To make causal inferences and conclude causal relationships from samples with observed variables, we must estimate the outcomes that are not factual. Some general assumptions are needed under the potential outcome framework.

- **SUTVA (Stable Unit Treatment Value Assumption):** The potential outcomes of one unit are not affected by the treatment status of other units and there is only one form or version of each treatment.
- **Consistency:** The potential outcome that corresponds to the treatment taken is actually the observed outcome:

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- **Exchangeability:** The potential outcomes are independent of the treatment status given all confounding variables:

$$(Y_i(1), Y_i(0)) \perp A_i | U_i$$

- **Positivity:** The conditional probability of taking different treatments given the confounding variables must be greater than zero and less than one:

$$0 < \Pr(A_i = a | U_i) < 1, \qquad where\ a = 0,1$$

Nonetheless, assumptions are not facts and they can be false. Sometimes we are not even able to test some assumptions from the knowledge at hand (Imbens & Rubin, 2015). If inferences are based on false assumptions, we cannot expect valid estimates.

2.2 Principal Stratification Framework

To identify and estimate the average causal effects in a certain population, we must ensure the comparison groups represent the same population. Instead of adjusting for the pretreatment variables, Balke and Pearl (1994) introduced the concept of "response variable" that substitutes the multi-dimensional pretreatment variables and divides the population into equivalent classes so that treatment groups within the same class are comparable. Frangakis and Rubin (2002) conceptualized the term "equivalence classes" as "principal strata" and proposed the principal stratification framework particularly targeting at post-treatment variables. Within their framework, units are cross-classified into different strata "based on their joint potential values of the post-treatment variable under each of the treatments being compared" without introducing post-treatment selection bias. The treatment effects within each stratum are thus indeed causal effects.

Applications using principal stratification abound in the literature. Consensus has been reached on the success of this framework for nonadherence issues in RCTs (Pearl, 2011; VanderWeele, 2011). For a placebo-controlled RCT with binary treatments and binary adherence, using notations in section 2.1, the population can be partitioned into four strata based on the pair of potential treatment taken under different treatment assignment, $(A_i(1), A_i(0))$, as shown in Table 2.1.

Table 2.1 Principal strata classified by A(0) and A(1) for placebo-controlled RCTs.

		Z=0	
		A(0)=0	A(0)=1
Z=1	A(1)=0	Never-taker	Defier
	A(1)=1	Complier	Always-taker

For those that never take the treatment whichever group they are assigned, i.e., ($A_i(1) = 0, A_i(0) = 0$), they are defined as *never-takers*. Those that always take the treatment whichever group they are assigned, i.e., ($A_i(1) = 1, A_i(0) = 1$), are referred to as *always-takers*. *Compliers* are defined as those who take the treatment if randomized to treatment and take the placebo if randomized to placebo ($A_i(1) = 1, A_i(0) = 0$), and *defiers* are those always do the opposite to what they are assigned ($A_i(1) = 0, A_i(0) = 1$). The subpopulations of never-takers, always-takers and defiers are referred to as *noncompliers*. After the classification, we can easily define any causal effect within a specific stratum. For example, the treatment effect in the complier stratum is called *local average treatment effect (LATE)* or *complier average causal effect (CACE).* As the principal strata are defined by the potential outcomes of the post-treatment variable and we can never observe all the potential outcomes simultaneously, it is impossible for us to know exactly which stratum a unit belongs to. Nonetheless, it is still critical to use principal stratification to define causal effects because it helps us clarify which causal quantity we can identify and it is the foundation for subsequent estimation and interpretation.

2.3 Identification of Causal Estimands in RCTs with Treatment Nonadherence

An *estimand* is "the target of estimation for a particular trial objective". It asks the question "what is to be estimated?" (ICH E9 working group, 2020). Thus, an estimand can be of any form even

without a practical interpretation as long as it aligns with the scientific goals. In addition to corresponding to the research objectives, a *causal estimand* should also provide a reasonable causal interpretation. A fundamental attribute of a causal estimand, especially when the interest lies in estimating causal effects, is that it targets the same group of units (Frangakis & Rubin, 2002). Researchers and scientists often use terms that involve causal implication such as "causation" or "causality" with extreme caution as it is usually easy to confuse "association" which can be an estimand, with "causation" which must be a causal estimand if it is the study objective.

In a RCT, the ITT estimand is a causal estimand because randomization ensures the comparison groups represent the same population. However, with the existence of treatment nonadherence, the ITT estimand fails to answer the correct research question if the causal effect of treatment rather than of assignment is of interest. The traditional per-protocol estimand and as-treated estimand, on the other hand, take into account treatment nonadherence. By definition, the traditional per-protocol estimand only includes those who are "observed to follow their treatment assignment" and the as-treated estimand ignores treatment assignment and directly compares the actual treatments received (Shrier et al., 2014). These two estimands describe some particular treatment effects. However, we are hesitant to conclude they are causal estimands because we have limited information to envision what the participants in one group would have done if they had been in the other group. Specifically, since participants are no longer randomized, the treatment groups are not guaranteed to be comparable for the per-protocol estimand and as-treated estimand. Hence, these two estimands will not be included in our subsequent discussion.

We consider the simplest way to measure compliance in a RCT which is to dichotomize it as whether participants comply with their assigned treatment or not. For a placebo-controlled RCT with binary treatments and binary compliance, Angrist et al. (1996) point-identified the CACE

using instrumental variables (IV) without assuming any parametric models. Their identification was based on two critical assumptions: the exclusion restriction (ER) assumption and the monotonicity assumption. The ER assumption assumes that there is no direct effect of treatment assignment Z on outcome Y. In other words, the effect of Z on Y can only go through the effect of treatment A on Y. The monotonicity assumption requires that assignment Z influences treatment A monotonically, i.e., there are no defiers (Imbens & Angrist, 1994). Under these two assumptions, Angrist et al. (1996) identified the CACE as the ratio of the ITT causal effect of Z on Y and the ITT causal effect of Z on A. Using the same logic of partitioning the population into four principal strata (Table 2.1), Imbens and Rubin (1997) proposed a Bayesian approach to infer the causal estimands of CACE and defier average causal effect (DACE) using EM and data augmentation algorithms for posterior maximum likelihood estimation. Their approach can relax the ER and monotonicity assumptions from the nonparametric IV approach. However, Imbens and Rubin (1997) also found that estimation for CACE can be more accurate compared to the nonparametric method if under these two assumptions. Both Little and Yau (1998) and Little et al. (2009) compared the traditional IV approach and the maximum likelihood (ML) estimation for CACE and argued that the model-based estimator was more efficient than the IV estimator, but was sensitive to model specification and required large sample size.

The identification and estimation of CACE has also been extensively studied in the presence of covariates. With covariates, some assumptions for identifying CACE can be satisfied, or removed. Hirano et al. (2000) extended the work of Imbens & Rubin (1997) by incorporating the covariates. They formulated separate ER assumptions for subgroups given covariates and examined the sensitivity to violations of different ERs. They claimed a weakly identified model without ERs which needed a more careful selection of likelihood function and prior distribution.

Instead of relying on choosing proper priors and likelihood functions, Jo (2002) explored the identifiability of CACE by relying on limited covariate information. Specifically, Jo (2002) modified the Bayesian approach of Imbens & Rubin (1997) by relaxing the ER assumption and adding two additional functional assumptions which were believed to be more reasonable and applicable in practice. Frangakis et al. (2002) expanded the identification problem associated with noncompliance to clustered data and proposed a Bayesian framework for the *clustered encouragement design*. Further, they used covariates to relax the ER assumption and predicted the principal strata defined by the potential compliance.

Hitherto, the discussion only involved placebo-controlled RCTs. Attention has been paid to RCTs with nonadherence beyond placebo-controlled RCTs as well. Roy et al. (2008) considered the identification of CACE in the setting with two active treatments. Without covariate information, they partially identified CACE by setting bounds under two more assumptions, one of which is the treatment access restriction assumption which disallows subjects in one treatment group to access the treatment in the other group. The other one is the modified monotonicity assumption which assumes that the probability of compliance of one treatment assignment group is higher for those who would comply to the other treatment assignment compared to those who would not. Based on the covariates, however, Roy et al. (2008) were able to model the marginal distributions of the compliance status and point-identify the CACE. Long et al. (2010) proposed a Bayesian approach following Imbens & Rubin (1997) for the RCTs with three treatment arms. Among the treatment arms, two are active treatments and the third one is the control arm. To reduce the number of principal strata, two extra assumptions were made as well. One of the assumptions is no access to the other treatment assumption as in Roy et al. (2008). The other one is a modified monotonicity assumption which assumes that subjects who comply with one

treatment would always comply with the other treatment. Under these assumptions, Long et al. (2010) modeled the principal compliance directly without incorporating covariates. Yuan et al. (2018) proposed a novel way to identify certain principal causal effects utilizing the features of a multisite design. They separated their control into two different types to identify their estimands of interest. Similarly, they imposed two structural assumptions to reduce the number of principal strata. Then they made a key assumption about the site-level zero correlation between the distribution of the principal strata and the principal causal effects. Under this strong assumption, they proved that the CACE can be identified. Yuan et al. (2018) also investigated the use of covariates to relax the above strong assumption and obtained a consistent estimator of CACE by conditioning on the covariates.

CHAPTER 3

A NONPARAMETRIC APPROACH

3.1 Study Setup and Causal Estimands

In this section, we first describe our study setup, that is, a RCT with two active treatments subject to nonadherence, under the Neyman-Rubin potential outcome and principal stratification frameworks. We extend notations introduced in section 2.1. Here, the treatment assignment Z_i takes values 1 and 2, with 1 for treatment 1 and 2 for treatment 2. There are therefore three possible values for the potential treatment received $A_i(z)$: 0, 1, and 2. Thus, given the three values of $A_i(z)$ under each treatment, we can define nine principal strata as shown in the 3×3 Table 3.1.

Table 3.1 Principal strata classified by A(1) and A(2) for RCTs with two active treatments.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1	2. Defier (DF) π_2	3. Partial-2-complier (P2C) π_3
	A(1)=1	4. Partial-1-complier (P1C) π_4	5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2	7. Defier (DF) π_7	8. Defier (DF) π_8	9. Always-2-taker (A2T) π_9

We number the table cells row by row and denote their corresponding probabilities as π_g, $g = 1, 2, \ldots, 9$. Analogous to the 2×2 Table 2.1, we define *never-takers* (NT) and *compliers* (COMP) in the same way as in section 2.2. The always-takers for the placebo-controlled RCTs have now been separated into the *always-1-takers* (A1T) who always take treatment 1 regardless of which treatment they are assigned, and the *always-2-takers* (A2T) who always take treatment 2

regardless of which treatment they are assigned. We name those who comply with only one treatment but do not take anything if assigned to the other treatment as *partial-compliers* (PC). Among the partial compliers, we further differentiate those who only comply with treatment 1 but take nothing if assigned to treatment 2 as *partial-1-compliers* (P1C) and those vice versa as *partial-2-compliers* (P2C). Finally, we call the subpopulations in the three remaining strata *defiers* (DF) as they do not show any systematically consistent preference but show contradictory actions to their own decisions. For example, for those with $(A_i(1) = 2, A_i(2) = 0)$, if assigned to treatment 1 they choose to take treatment 2, but if assigned to treatment 2 they choose to taking nothing. It seems that they ordered their preference as nothing>treatment 2>treatment 1. However, if this is the case, they would take nothing if assigned to treatment 1 at the beginning. We further subdivide *defiers* into two categories. Based on their similarity in decision making for the treatment received, for those with $(A_i(1) = 0, A_i(2) = 1)$ or $(A_i(1) = 2, A_i(2) = 0)$, they are named as *irrationalists* (IR). For those with $(A_i(1) = 2, A_i(2) = 1)$, they are *flip-floppers* (FF).

Under the above setting, we next define our causal estimand of interest. For individual i, let $Y_i(z, A_i(z))$ be the potential outcome if the individual is randomly assigned to treatment z with $z = 1$ or 2. This two-index potential outcome can be simplified to the one-index potential outcome $Y_i(z)$ as they both depend solely on the values of z. However, the notation $Y_i(z)$ might be confused with our previously defined potential outcome of treatment received $Y_i(a)$. Thus, we will use $Y_i(z, A_i(z))$ to avoid confusion in subsequent sections and the one-index potential outcome will always represent the potential outcome under different treatment values a. It is worth mentioning that both $Y_i(z, A_i(z))$ and $Y_i(a)$ depends merely on the values of one variable, which is z or a. There is another potential outcome $Y_i(z, a)$ that depends on both values of z and a. For example,

if the treatment assignment has a direct effect on the outcome, both variables Z_i and A_i will have an impact on Y_i and the corresponding potential outcome will be $Y_i(z, a)$.

We define the ITT effect comparing treatment assignment 2 with treatment assignment 1 as $Y_i\big(2, A_i(2)\big) - Y_i(1, A_i(1))$ at the individual level and as $E[Y_i\big(2, A_i(2)\big) - Y_i(1, A_i(1))]$ at the population level. With nonadherence in our setting and aiming at identifying the causal effects of treatment received, the ITT estimand is clearly not the one we pursue. The average treatment effect (ATE) estimand compares the potential outcomes of treatment 2 and treatment 1, i.e., $E[Y_i(2) - Y_i(1)]$ for the entire population. However, ATE is not readily identified without the strong constant effect assumption which assumes the causal effects are the same across individuals. Alternatively, we identify the average treatment effect for a specific principal stratum, the compliers, that is, we identify the average treatment effect for those who will always comply to the treatments to which they are randomly assigned. Thus, in our study setting, we aim at identifying the complier average causal effect, CACE.

3.2 Identification Strategies of the Causal Estimand

In section 2.1, we reviewed several general assumptions to identify the causal effects. In this section, we carefully examine extensions of these assumptions to our study setting. First, we modify the consistency assumption that connects the observed variables with the potential variables. As we assign values 1 and 2 to our two treatments, the relationship between the observed A_i and the potential $A_i(z)$ is then $A_i = A_i(1)(2 - Z_i) + A_i(2)(Z_i - 1)$. Similarly, the relationship between the observed Y_i and the potential $Y_i(a)$ becomes $Y_i = A_i(2 - A_i)Y_i(1) + \frac{1}{2}A_i(A_i - 1)Y_i(2) + \frac{1}{2}(A_i - 1)(A_i - 2)Y_i(0)$. Second, we modify the exchangeability assumption to include the three potential outcomes, that is, $\big(Y_i(0), Y_i(1), Y_i(2)\big) \perp A_i \mid U_i$. To find the

identification strategies for our causal estimand of interest, we make some additional assumptions as follows.

Assumption I (Exclusion restriction, ER): There is no direct effect from Z to Y, i.e., the effect of Z on Y can only go through the effect of A on Y, or $Y_i(z, A_i(z)) = Y_i(A_i(z))$ for $z = 1$ or 2.

Assumption II (Monotonicity or No defiers): There are no individuals with $(A_i(1) = 0, A_i(2) = 1)$, $(A_i(1) = 2, A_i(2) = 0)$, or $(A_i(1) = 2, A_i(2) = 2)$, i.e., $\pi_2 = \pi_7 = \pi_8 = 0$.

Assumption II.a (No irrationalists): There are no individuals with $(A_i(1) = 0, A_i(2) = 1)$ or $(A_i(1) = 2, A_i(2) = 0)$, i.e., $\pi_2 = \pi_7 = 0$.

Assumption II.b (No flip-floppers): There are no individuals with $(A_i(1) = 2, A_i(2) = 1)$, i.e., $\pi_8 = 0$.

The monotonicity assumption is plausible as we assume that when participants consent to enter a trial, it is less likely for them to completely disobey the treatment implementation procedure without showing any reasonable logic. We also separate the assumption into two sub-assumptions for our sensitivity discussion later.

Assumption III (No partial-compliers): There are no individuals with $(A_i(1) = 0, A_i(2) = 2)$ or $(A_i(1) = 1, A_i(2) = 0)$, i.e., $\pi_3 = \pi_4 = 0$.

This assumption states that if participants choose to comply with one treatment, it is implausible that they choose to take nothing when assigned to the other treatment. It is more specific to the study environment. There is a possibility that it can be violated and it is not straightforward to justify it. We will discuss deviation from this assumption in section 3.3.3.

With the above assumptions, we now formulate the causal effect of A on Y. First, under the SUTVA, consistency and exclusion restriction assumptions, we establish a relationship

between the ITT effect of Z on Y, i.e., $Y_i\big(2, A_i(2)\big) - Y_i(1, A_i(1))$ and the causal effect of A on Y at the individual level. That is,

$$
\begin{aligned}
& Y_i(2, A_i(2)) - Y_i(1, A_i(1)) \\
&= Y_i(A_i(2)) - Y_i(A_i(1)) \\
&= [A_i(2)\big(2 - A_i(2)\big)Y_i(1) + \tfrac{1}{2}A_i(2)(A_i(2) - 1)Y_i(2) + \tfrac{1}{2}(A_i(2) - 1)(A_i(2) - 2)Y_i(0)] \\
&- [A_i(1)\big(2 - A_i(1)\big)Y_i(1) + \tfrac{1}{2}A_i(1)(A_i(1) - 1)Y_i(2) + \tfrac{1}{2}(A_i(1) - 1)(A_i(1) - 2)Y_i(0)] \\
&= -\tfrac{1}{2}\big(A_i(2) - A_i(1)\big)\big(Y_i(2) - Y_i(1)\big) + \tfrac{1}{2}\big(A_i(2) - A_i(1)\big)\big(A_i(2) + A_i(1)\big)\big(Y_i(2) - Y_i(1)\big) \\
&+ (A_i(2) - A_i(1))(\tfrac{3}{2} - A_i(2) - A_i(1))(Y_i(1) - Y_i(0)) \qquad (3.1)
\end{aligned}
$$

The first equality of equation (3.1) holds under Assumption I which reduces the two-index potential outcome $Y_i(z, A_i(z))$ to the one-index potential outcome $Y_i(A_i(z))$ for $z = 1$ or 2. The second equality follows the consistency assumption. Next, we can write down the average ITT effect by taking expectations to both the left- and right-hand sides of equation (3.1):

$$
\begin{aligned}
& E[Y_i(2, A_i(2)) - Y_i(1, A_i(1))] \\
&= -\tfrac{1}{2}E\big[\big(A_i(2) - A_i(1)\big)\big(Y_i(2) - Y_i(1)\big)\big] \\
&+ \tfrac{1}{2}E\big[\big(A_i(2) - A_i(1)\big)\big(A_i(2) + A_i(1)\big)\big(Y_i(2) + Y_i(1)\big)\big] \\
&+ E[(A_i(2) - A_i(1))(\tfrac{3}{2} - A_i(2) - A_i(1))(Y_i(1) - Y_i(0))] \\
&= [\tfrac{1}{2}E(Y_i(2) - Y_i(1)|A_i(2) - A_i(1) = -1) - \tfrac{1}{2}E(Y_i(2) - Y_i(0)|A_i(2) - A_i(1) = -1) \\
&- \tfrac{1}{2}E(Y_i(1) - Y_i(0)|A_i(2) - A_i(1) = -1)]Pr(A_i(2) - A_i(1) = -1) \\
&+ [-E(Y(2) - Y(1)|A(2) - A(1) = 2) + 2E(Y(2) - Y(0)|A(2) - A(1) = 2) \\
&- E(Y(1) - Y(0)|A(2) - A(1) = 2)]Pr(A(2) - A(1) = 2) \\
&+ [-\tfrac{1}{2}E(Y_i(2) - Y_i(1)|A_i(2) - A_i(1) = 1) + \tfrac{3}{2}E(Y_i(2) - Y_i(0)|A_i(2) - A_i(1) = 1)
\end{aligned}
$$

$$- \frac{3}{2}E(Y_i(1) - Y_i(0)|A_i(2) - A_i(1) = 1)]Pr(A_i(2) - A_i(1) = 1)$$

$$= E[Y_i(0) - Y_i(1)|A_i(2) - A_i(1) = -1]\, Pr(A_i(2) - A_i(1) = -1)$$

$$+ E[Y_i(2) - Y_i(0)|A_i(2) - A_i(1) = 2]\, Pr(A_i(2) - A_i(1) = 2)$$

$$+ E[Y_i(2) - Y_i(1)|A_i(2) - A_i(1) = 1]Pr(A_i(2) - A_i(1) = 1)$$

$$= E[Y_i(2) - Y_i(1)|A_i(2) - A_i(1) = 1]Pr(A_i(2) - A_i(1) = 1) \quad (3.2)$$

The second equality in equation (3.2) holds under Assumption II and the last equality is true under Assumption III. Thus, with the two structural assumptions: Assumption II and III, we can reduce the nine principal strata to four, as shown in Table 3.2. Notice that, the right-hand side of equation (3.2) is a product of the complier average causal effect and the probability of $A_i(2) - A_i(1) = 1$. This probability is actually the average ITT effect of Z on A, i.e., $E[A_i(2) - A_i(1)]$, because $A_i(2) - A_i(1) = 0$ for NT, A1T and A2T. Therefore, we combine our findings into the following proposition.

Table 3.2 Principal strata classified by A(1) and A(2) for RCTs with two active treatments under Assumption II, the monotonicity assumption, and Assumption III, the no partial-compliers assumption.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1		
	A(1)=1		5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2			9. Always-2-taker (A2T) π_9

Proposition I: For a RCT with two active treatments subject to nonadherence, the complier average causal effect (CACE) can be identified under Assumption I, II, and III as the ratio of the ITT effect of Z on Y to the ITT effect of Z on A:

$$CACE = E[Y_i(2) - Y_i(1)|A_i(2) - A_i(1) = 1] = \frac{E[Y_i(2, A_i(2)) - Y_i(1, A_i(1))]}{E[A_i(2) - A_i(1)]}$$

To estimate CACE under Proposition I, we need therefore to estimate both the ITT effects of Z on Y and Z on A. Due to randomization and consistency, the ITT effect of Z on Y, $E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right]$ will be equal to $E(Y_i|Z_i = 2) - E(Y_i|Z_i = 1)$. Then, we can use the sample average to unbiasedly estimate $E(Y_i|Z_i = z), z = 1, 2$, i.e.,

$$\widehat{ITT}_y = \frac{\sum_{i\epsilon\{i:Z_i=2\}} Y_i}{N_2} - \frac{\sum_{i\epsilon\{i:Z_i=1\}} Y_i}{N_1}$$

where N_z is the number of individuals that are assigned to treatment z for $z = 1, 2$. Similarly, the unbiased estimator for the ITT effect of Z on A is

$$\widehat{ITT}_a = \frac{\sum_{i\epsilon\{i:Z_i=2\}} A_i}{N_2} - \frac{\sum_{i\epsilon\{i:Z_i=1\}} A_i}{N_1}$$

Thus, the estimated CACE is

$$\widehat{CACE} = \frac{\widehat{ITT}_y}{\widehat{ITT}_a} = \frac{\frac{\sum_{i\epsilon\{i:Z_i=2\}} Y_i}{N_2} - \frac{\sum_{i\epsilon\{i:Z_i=1\}} Y_i}{N_1}}{\frac{\sum_{i\epsilon\{i:Z_i=2\}} A_i}{N_2} - \frac{\sum_{i\epsilon\{i:Z_i=1\}} A_i}{N_1}} \quad (3.3)$$

3.3 Sensitivity of the Causal Estimand to Key Assumptions

As our identification strategies for CACE rely on several assumptions, in this section, we discuss how violations of key assumptions affect our causal estimand of interest. In other words, when these assumptions are violated, what causal effect will we be estimating using equation (3.3)? We consider deviations from these assumptions one at a time.

3.3.1 Violation of Assumption I (Exclusion Restriction)

Under the ER assumption, there will be no direct effect of treatment assignments on the outcomes. However, in some situations, the randomization may have an impact on the outcomes by stimulating other paths that go directly to the outcomes without passing through the treatment received. For example, in a behavioral trial, knowing which treatment arms they are assigned may change the participants' lifestyle, e.g., their smoking status, their workout plan, etc., and subsequently affect the health outcomes of interest. Thus, we assess the sensitivity of our causal estimand to the violation of the ER assumption while maintaining all the other assumptions. We rewrite the average ITT effect without the ER assumption as

$$
\begin{aligned}
& E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right] \\
&= E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] Pr(COMP) \\
&+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| NT\right] Pr(NT) \\
&+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| A1T\right] Pr(A1T) \\
&+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| A2T\right] Pr(A2T) \qquad (3.4)
\end{aligned}
$$

We let the direct effect of Z on Y be $H_i = Y_i(2, a) - Y_i(1, a)$ and the direct effect of A on Y be $G_i = Y_i(z, 2) - Y_i(z, 1)$. Then, for compliers, if we assume an additive effect of Z and A on Y (Angrist et al., 1996), equation (3.4) can be rewritten as

$$
\begin{aligned}
& \frac{E\left[Y_i\left(2, A_i(2)\right) - Y\left(1, A_i(1)\right)\right]}{E[A_i(2) - A_i(1)]} \\
&= \frac{E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right]}{Pr(COMP)} \\
&= E[G + H|COMP] + E[H|NT]\frac{Pr(NT)}{Pr(COMP)} + E[H|A1T]\frac{Pr(A1T)}{Pr(COMP)} + E[H|A2T]\frac{Pr(A2T)}{Pr(COMP)} \\
&= E[G|COMP] + E(H|COMP)\frac{Pr(COMP)}{Pr(COMP)} + E[H|NT]\frac{Pr(NT)}{Pr(COMP)}
\end{aligned}
$$

$$+ E[H|A1T]\frac{Pr(A1T)}{Pr(COMP)} + E[H|A2T]\frac{Pr(A2T)}{Pr(COMP)}$$

$$= CACE + \frac{E(H)}{Pr(COMP)} \quad (3.5)$$

Note that the CACE in equation (3.5) is defined as follows which is different from the CACE when the ER assumption is satisfied.

$$\begin{aligned} CACE = E[G|COMP] &= E[Y_i(z,2) - Y_i(z,1)|COMP] \\ &= E[Y_i(1,2) - Y_i(1,1)] \\ &= E[Y_i(2,2) - Y_i(2,1)] \end{aligned}$$

The ITT effect of Z on A will remain equal to the probability of compliers because there is no change of the population composition when the ER assumption is violated. In other words, the population still contains the four subpopulations that are from strata 1, 5, 6, and 9. Either increasing the proportion of compliers or decreasing the direct effects of treatment assignments on the outcomes will reduce the bias, as shown in equation (3.5).

3.3.2 Violation of Assumption II (Monotonicity)

Two of the assumptions (II and III) needed to identify our causal estimand of interest are structural assumptions. Deviation from these assumptions will result in a change of the population composition. In this section, we investigate the violation of the first structural assumption, i.e., monotonicity. Specifically, we will look at violation of the no irrationalists assumption (II.a) which will add principal strata 2 and 7 into our population and violation of the no flip-floppers assumption (II.b) which will add principal stratum 8 into our population.

3.3.2.1 Violation of Assumption II.a (No Irrationalists)

As the structure of the principal strata has changed (Table 3.3), the ITT effect of Z on A is no longer equal to the proportion of compliers. For irrationalists, we refer to those with $(A(1) = 0, A(2) = 1)$ as irrationalists 1 (IR1) and those with $(A(1) = 2, A(2) = 0)$ as irrationalists 2 (IR2) for the subsequent formulation.

Table 3.3 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption II.a, the no irrationalists assumption, is violated.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1	2. Irrationalist 1 (IR1) π_2	
	A(1)=1		5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2	7. Irrationalist 2 (IR2) π_7		9. Always-2-taker (A2T) π_9

The ITT effect of Z on A becomes

$$E[A_i(2) - A_i(1)] = (1-0) \times Pr(IR1) + (0-2) \times Pr(IR2) + (2-1) \times Pr(COMP)$$

$$= Pr(IR1) - 2Pr(IR2) + Pr(COMP)$$

$$= Pr(COMP) - [2Pr(IR2) - Pr(IR1)]$$

Then, our nonparametric estimator becomes

$$\frac{E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right]}{E[A_i(2) - A_i(1)]}$$

$$= E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| IR1\right] \frac{Pr(IR1)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]}$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| IR2\right] \frac{Pr(IR2)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]}$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] \left[1 + \frac{2Pr(IR2) - Pr(IR1)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]}\right]$$

$$= E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| IR\right] \frac{Pr(IR)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]}$$

$$+ E[Y_i(2) - Y_i(1) | COMP]$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] \frac{2Pr(IR2) - Pr(IR1)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]}$$

$$= CACE$$

$$+ \frac{Pr(IR)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]} E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| IR\right]$$

$$+ \frac{2Pr(IR2) - Pr(IR1)}{Pr(COMP) - [2Pr(IR2) - Pr(IR1)]} E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] \quad (3.6)$$

The last two items of equation (3.6) consist of the bias which relates to both the principal strata proportions and the principal strata causal effects. Suppose the participants in the IR stata are symmetric, i.e., the proportions of IR1 and IR2 are the same. Then, under this symmetry in principal strata assumption for irrationalists, the bias from equation (3.6) can be simplified to

$$Bias = \frac{Pr(IR1)}{Pr(COMP) - Pr(IR1)} \times$$

$$\{2E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| IR\right] + E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right]\} \quad (3.7)$$

The bias formula in equation (3.7) explicitly shows that bias from violation of the no-irrationalists assumption will be inversely proportional to the relative proportion of compliers to irrationalists. Equation (3.7) also shows that the bias will always exist even if the ITT effect of irrationalists is zero as long as there are irrationalists in the population unless the ITT effect of irrationalists is exactly half of the opposite of the complier average causal effect (CACE).

3.3.2.2 Violation of Assumption II.b (No Flip-Floppers)

The principal strata of the population after violating the no flip-floppers assumption are shown in Table 3.4. Similar as above, the ITT effect of Z on A is

$$E[A_i(2) - A_i(1)] = (2-1) \times Pr(COMP) + (1-2) \times Pr(FF) = Pr(COMP) - Pr(FF)$$

Thus, the nonparametric estimator becomes

$$\frac{E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right]}{E[A_i(2) - A_i(1)]}$$

$$= E[Y_i(2) - Y_i(1)|COMP] \frac{Pr(COMP)}{Pr(COMP) - Pr(FF)} - E[Y_i(2) - Y_i(1)|FF] \frac{Pr(FF)}{Pr(COMP) - Pr(FF)}$$

$$= CACE + E[Y_i(2) - Y_i(1)|COMP] \frac{Pr(FF)}{Pr(COMP) - Pr(FF)} - E[Y_i(2) - Y_i(1)|FF] \frac{Pr(FF)}{Pr(COMP) - Pr(FF)}$$

$$= CACE + \frac{Pr(FF)}{Pr(COMP) - Pr(FF)} \{E[Y_i(2) - Y_i(1)|COMP] - E[Y_i(2) - Y_i(1)|FF]\} \quad (3.8)$$

Table 3.4 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption II.b, the no flip-floppers assumption, is violated.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1		
	A(1)=1		5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2		8. Flip-floppers (FF) π_8	9. Always-2-taker (A2T) π_9

The second item of equation (3.8) forms the bias formula when the no flip-floppers assumption is violated. Similar implications can be found as for the irrationalists. Namely, the bias can be reduced if the relative proportion of compliers to flip-floppers increases. Only if there is no

variation in the causal effects of the treatment received for compliers and flip-floppers, that is, the causal effects of treatment received is constant across the principal strata, the bias will be zero.

3.3.3 Violation of Assumption III (No Partial-Complier)

We discuss the sensitivity of our causal estimand to the violation of the second structural assumption: no partial-compliers in this section. The structure of the principal strata is changed to Table 3.5. The ITT effect of Z on A is

$$E[A_i(2) - A_i(1)] = (2-0) \times Pr(P2C) + (0-1) \times Pr(P1C) + (2-1) \times \Pr(COMP)$$
$$= 2Pr(P2C) - Pr(P1C) + Pr(COMP)$$

The nonparametric estimator then becomes

$$\frac{E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right)\right]}{E[A_i(2) - A_i(1)]}$$

$$= E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| P2C\right] \frac{Pr(P2C)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)}$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| P1C\right] \frac{Pr(P1C)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)}$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] \left[1 - \frac{2\,Pr(P2C) - Pr(P1C)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)}\right]$$

$$= E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| PC\right] \frac{Pr(PC)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)}$$

$$+ E[Y_i(2) - Y_i(1) | COMP]$$

$$+ E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| COMP\right] \frac{Pr(P1C) - 2Pr(P2C)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)}$$

$$= CACE$$

$$+ \frac{Pr(PC)}{2Pr(P2C) - Pr(P1C) + Pr(COMP)} E\left[Y_i\left(2, A_i(2)\right) - Y_i\left(1, A_i(1)\right) \middle| PC\right]$$

$$+\frac{Pr(P1C)-2Pr(P2C)}{2Pr(P2C)-Pr(P1C)+Pr(COMP)}E\left[Y_i\left(2,A_i(2)\right)-\ Y_i\left(1,A_i(1)\right)\middle|COMP\right] \quad (3.9)$$

We can derive the bias from the last two items of equation (3.9). Similarly, under the symmetry in principal strata assumption for partial-compliers, the proportion of partial-1-compliers is identical to that of partial-2-compliers, the bias can be simplified to

$$Bias=\frac{Pr(P1C)}{Pr(P1C)+Pr(COMP)}\times$$

$$\left\{2E\left[Y_i\left(2,A_i(2)\right)-\ Y_i\left(1,A_i(1)\right)\middle|PC\right]-E\left[Y_i\left(2,A_i(2)\right)-\ Y_i\left(1,A_i(1)\right)\middle|COMP\right]\right\} \quad (3.10)$$

Again, the bias will decrease as the relative proportion of compliers to partial-compliers increase. Even when the ITT effect of partial-compliers is zero, the estimator is still biased. Only if the ITT effect of partial-compliers is on average half of the complier average causal effect will the bias be zero.

Table 3.5 Principal strata classified by A(1) and A(2) for RCTs with two active treatments when Assumption III, the no partial-compliers assumption, is violated.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1		3. Partial-2-complier (P2C) π_3
	A(1)=1	4. Partial-1-complier (P1C) π_4	5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2			9. Always-2-taker (A2T) π_9

3.4 Simulation Studies

3.4.1 Simulation Setup

In this section, we introduce a series of simulations to evaluate the performance and sensitivity of our nonparametric estimator in equation (3.3). The data generating process follows 5 steps:

1) Generate the treatment assignment. We use a Bernoulli distribution with a probability of 0.5 to generate the treatment assignment Z_i for individual i.

2) Generate the nine principal strata. We use a multinomial distribution with nine pre-specified probabilities $\pi_1, \pi_2, \dots, \pi_9$, where $\sum_{g=1}^{9} \pi_g = 1$ to generate the principal strata G_i for individual i.

3) Obtain the observed treatment received. Given the principal strata G_i generated in step 2), we can obtain the values of the potential treatment received $(A_i(1), A_i(2))$ for individual i. Then, with the known treatment assignment Z_i, the observed treatment received A_i will be equal to $A_i(1)$ if $Z_i = 1$ and $A_i(2)$ if $Z_i = 2$.

4) Generate the potential outcomes. We consider both the continuous outcomes and the binary outcomes. We generate the potential outcomes $Y_i\big(1, A_i(1)\big)$ and $Y_i(2, A_i(2))$ separately.

a. For the continuous potential outcomes,

$$Y_i\big(1, A_i(1)\big) = \alpha_0 + \alpha_1 g_{1i} + \alpha_2 g_{2i} + \alpha_3 g_{3i} + \alpha_4 g_{4i} + \alpha_5 g_{5i} + \alpha_6 g_{6i} + \alpha_7 g_{7i} + \alpha_8 g_{8i} + \alpha_9 g_{9i} + \delta_{1z} I_{zi} + \epsilon_i$$

$$Y_i\big(2, A_i(2)\big) = \beta_0 + \beta_1 g_{1i} + \beta_2 g_{2i} + \beta_3 g_{3i} + \beta_4 g_{4i} + \beta_5 g_{5i} + \beta_6 g_{6i} + \beta_7 g_{7i} + \beta_8 g_{8i} + \beta_9 g_{9i} + \delta_{2z} I_{zi} + \epsilon_i$$

where g_{ji} is an indicator variable with $g_{ji} = 1$ if individual i is in stratum j and 0 otherwise. I_{zi} indicates the treatment assignment with $I_{zi} = 1$ if individual i is assigned to treatment 1 and 0 if assigned to treatment 2. The one error term ϵ_i, shared by the two counterfactuals, follows a standard normal distribution $N(0, 1)$.

b. For the binary potential outcomes,

$$logit\big(\Pr(Y_i(1, A_i(1)) = 1)\big) = \alpha_0 + \alpha_1 g_{1i} + \alpha_2 g_{2i} + \alpha_3 g_{3i} + \alpha_4 g_{4i} + \alpha_5 g_{5i} + \alpha_6 g_{6i} + \alpha_7 g_{7i} + \alpha_8 g_{8i} + \alpha_9 g_{9i} + \delta_{1z} I_{zi}$$

$$logit(\Pr(Y_i(2, A_i(2)) = 1)) = \beta_0 + \beta_1 g_{1i} + \beta_2 g_{2i} + \beta_3 g_{3i} + \beta_4 g_{4i} + \beta_5 g_{5i} + \beta_6 g_{6i} + \beta_7 g_{7i} + \beta_8 g_{8i} + \beta_9 g_{9i} + \delta_{2z} I_{zi}$$

where the definition of the notations are the same as the continuous case.

5) Obtain the observed outcome. Given the treatment assignment Z_i and the potential outcomes $Y_i(z, A_i(z))$, $z = 1, 2$ generated in step 4), the observed outcome Y_i will be $Y_i\big(1, A_i(1)\big)$ if $Z_i = 1$ and $Y_i\big(2, A_i(2)\big)$ if $Z_i = 2$.

After all data are generated, we can then obtain the estimate of our causal estimand of interest from the sample quantities using equation (3.3). We run five sets of simulations for the five scenarios A to E for both the continuous outcomes and the binary outcomes. Scenario A is to evaluate the performance of the nonparametric estimator when all assumptions are met. Scenarios B to E are the sensitivity analysis when one of the assumptions described in section 3.2 is violated. We summarize the parameters used to generate the potential outcomes for each scenario in Table 3.6. The true value of CACE is set to be 6 for the continuous case and is around 0.35 for the binary case. To simplify exposition, we call the irrationalists, flip-floppers, and partial-compliers violators as their existence is due to violations of a certain assumption. For each scenario, we vary the principal strata proportion of compliers from 50% to 75% with an increment of 5%. On the other hand, for scenarios B to E, we fix the total principal strata proportions for never-takers, always-1-takers, and always-2-takers to be 20%. Then, the principal strata proportions of the violators will vary from 30% to 5% corresponding to the varying principal strata proportion of compliers. As a result, the relative proportions of compliers to violators will be increasing as the principal strata

proportion of compliers is increasing and the principal strata proportions of the violators are decreasing.

We consider various sample sizes with $n = 500, 1000$ or 2000. As there are six principal strata proportions, three sample sizes, two types of outcomes and five scenarios, we will therefore run $6 \times 3 \times 2 \times 5 = 180$ sets of simulations in total each with 10000 iterations. In addition, within each iteration of the simulation, we draw 2000 bootstrap samples to calculate the bootstrap standard errors of our nonparametric estimator.

Table 3.6 Summary of the parameter values used to generate the potential outcomes for the nonparametric approach. Scenario A is when all assumptions are satisfied. Scenario B is when Assumption I, the exclusion restriction assumption, is violated. Scenario C is when Assumption II.a, the no irrationalists assumption, is violated. Scenario D is when Assumption II.b, the no flip-floppers assumption, is violated. Scenario E is when Assumption III, the no partial-compliers assumption, is violated.

	Continuous Outcome					Binary Outcome				
Scenario	A	B	C	D	E	A	B	C	D	E
α_0	0	0	0	0	0	0	0	0	0	0
α_1	-3	-3	-3	-3	-3	0.5	0.5	0.5	0.5	0.5
α_2	0	0	-1	0	0	0	0	-2	0	0
α_3	0	0	0	0	-1	0	0	0	0	-3
α_4	0	0	0	0	7	0	0	0	0	0.2
α_5	8	8	8	8	8	0.5	0.5	0.5	0.5	0.5
α_6	4	4	4	4	4	-0.5	-0.5	-0.5	-0.5	-0.5
α_7	0	0	5	0	0	0	0	0.3	0	0
α_8	0	0	0	6	0	0	0	0	0.2	0
α_9	12	12	12	12	12	0.5	0.5	0.5	0.5	0.5
β_0	0	0	0	0	0	0	0	0	0	0
β_1	-3	-3	-3	-3	-3	0.5	0.5	0.5	0.5	0.5
β_2	0	0	3	0	0	0	0	-0.2	0	0
β_3	0	0	0	0	10	0	0	0	0	0.3
β_4	0	0	0	0	-1	0	0	0	0	-2
β_5	8	8	8	8	8	0.5	0.5	0.5	0.5	0.5
β_6	10	10	10	10	10	1	1	1	1	1
β_7	0	0	-4	0	0	0	0	-3	0	0
β_8	0	0	0	3	0	0	0	0	-0.8	0
β_9	12	12	12	12	12	0.5	0.5	0.5	0.5	0.5
δ_{1z}	0	-1	0	0	0	0	-0.5	0	0	0
δ_{2z}	0	-2	0	0	0	0	-0.3	0	0	0

3.4.2 Performance Metrics

We use the following metrics to evaluate the performance and sensitivity of our nonparametric estimator.

1. Mean bias: the average of the difference between the estimated CACE and the true CACE,

$$\frac{\sum_{i=1}^{10000}(\widehat{CACE}_i - CACE)}{10000}$$

2. Percent bias (%): the average of the bias relative to the true CACE,

$$\frac{\sum_{i=1}^{10000}(\widehat{CACE}_i - CACE)}{10000 \times CACE} \times 100$$

3. Mean standard error: the average of the standard errors from each iteration of the simulations.

$$\frac{\sum_{i=1}^{10000} SE(\widehat{CACE}_i)}{10000}$$

   where each SE($\widehat{CACE}$) is estimated by 2000 bootstrapped samples within a simulation.

4. Root mean squared error: the square root of the average of the squared bias,

$$\sqrt{\frac{\sum_{i=1}^{10000}\left(\widehat{CACE}_i - CACE\right)^2}{10000}}$$

5. Standard error ratio: the ratio of the mean standard error to the standard deviation of the 10000 estimated CACE

6. Coverage rate (%): the percent of the 95% confidence intervals that cover the true CACE. The 95% confidence interval is calculated as

$$(\widehat{CACE} - 1.96 \times SE, \widehat{CACE} + 1.96 \times SE)$$

3.4.3 Results

We present the results of scenarios A to E for the continuous outcomes here. Results for the binary outcomes show similar patterns and can be found in Appendix A.

3.4.3.1 Performance Analysis

The top two panels of Figure 3.1 report the mean bias and percent bias of the estimator. Our nonparametric method can yield slightly biased estimators for CACE as both the bias and percent bias approach to zero when $n = 500, 1000\ or\ 2000$. The estimators are less sensitive to the principal strata proportion of compliers when there is an adequate number of individuals in the sample. For example, when $n = 2000$, the estimators gather around the true CACE with approximately zero bias and percent bias for the various principal strata proportions of compliers. However, when $n = 500$, the bias and percent bias are more scattered across the principal strata proportions of compliers. This is consistent with the results of mean standard errors that we discuss below. In addition, this pattern is as expected because the size of compliers will always be larger in a sample with higher percentage of compliers than in a sample of the same sample size but with lower percentage of compliers. As we are estimating the average treatment effect in compliers, the larger size of compliers will therefore produce relatively smaller bias.

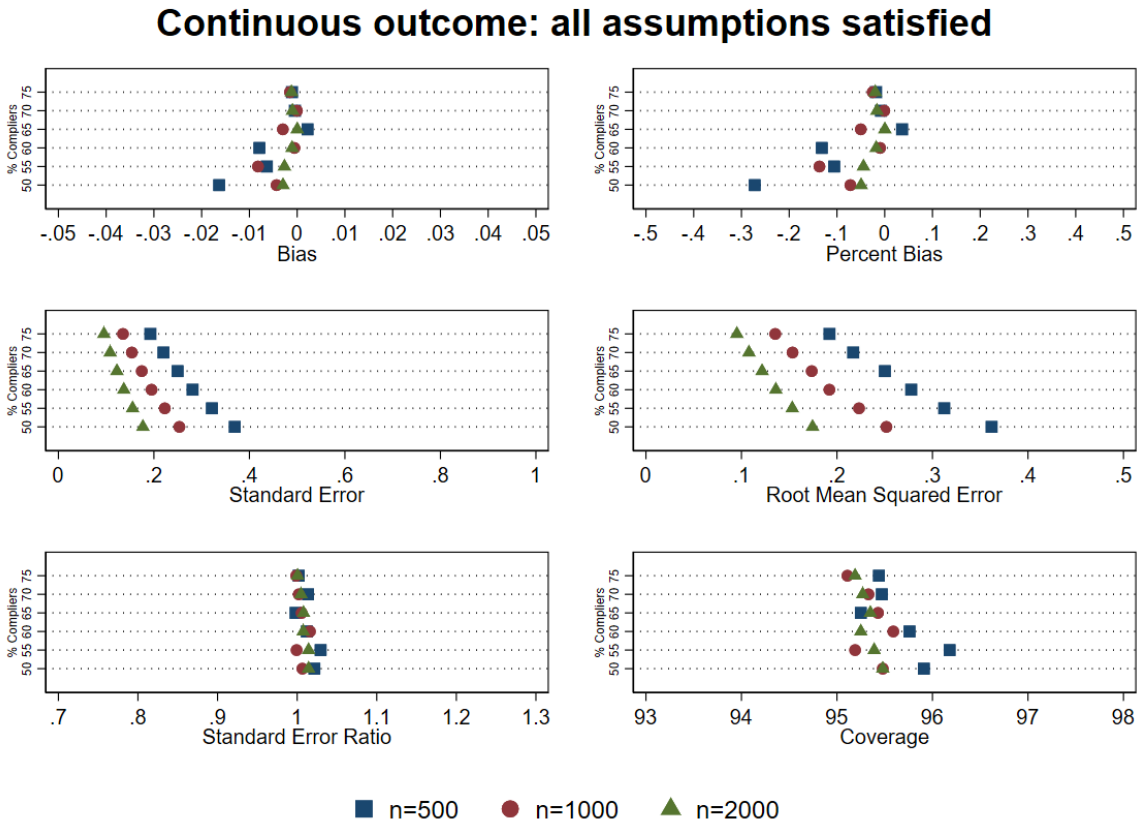

Figure 3.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the nonparametric estimator across proportions of compliers for the continuous outcome when all assumptions are satisfied (scenario A). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

Similar patterns are observed for the mean standard errors and root mean squared errors as shown in the middle two panels of Figure 3.1. The larger the sample size, the higher the principal strata proportion of compliers, the smaller the mean standard errors. We anticipate that the root mean squared errors will have the same pattern as the mean standard errors. That is, the larger the sample size, the larger the complier size, the smaller the root mean squared errors, the better the estimators.

We use the standard error ratios to compare the mean standard errors with the empirical standard errors of the estimator. The bottem left corner of Figure 3.1 shows that the standard error ratios are all roughly one, suggesting that the bootstrap standard errors of the nonparametric estimator can accurately estimate the empirical standard deviations of the estimated CACEs. The coverage rates, which are shown at the bottom right corner of Figure 3.1, are around 95% when $n = 1000\ or\ 2000$. However, they are inflated when $n = 500$, especially for samples with lower principal strata proportions of compliers.

3.4.3.2 Sensitivity Analysis

From the bias formula derived in section 3.3 when violations of Assumption I, II, and III occur, in general, the bias will depend on the principal strata proportions as well as certain treatment effects including the direct effects of treatment assignments on the outcomes and the principal strata ITT effects. However, we will only focus on evaluating the consequnces from changes of the principal strata proportions because 1) the direct impact from violations of the assumptions is a change of the population composion; and 2) either the direct effects of treatment assignments or the principal strata ITT effects can be thought as inherent properties of the principal strata which will be a constant within a principal stratum. As a result, we will not be interested in the magnitudes or

signs of the bias. We will, on the other hand, be interested in whether there will be some trends or patterns from the changes of the principal strata proportions.

We discuss the bias due to violation of ER separately from the bias due to the three types of violators because bias due to violation of ER is more sensitive to the specific values of the direct effects of treatment assignments. Although increasing the percentage of compliers in a sample will decrease the bias as shown in Figure 3.2, it will not approach to zero unless the direct effects of treatment assignments are closer to zero. However, as indicated in Figure 3.3, Figure 3.4, and Figure 3.5, the bias due to violations of the structural assumptions can be as close as zero if the relative proportion of compliers to violators is sufficiently large. Thus, we are able to minimize the bias due to the violators by maximizing the percentage of compliers and minimizing the percentage of violators.

Increasing sample size will not reduce the bias, but it will make the estimators more precise as shown in the standard error panels of Figure 3.2, Figure 3.3, Figure 3.4, and Figure 3.5. The standard errors will also be smaller when there are more compliers or less violators in the sample. The root mean squared errors show the performance of an estimator. Similar to the patterns of the bias and the mean standard errors, increasing the proportion of compliers or decreasing the proportions of violators will lower the root mean squared errors and make the estimators better.

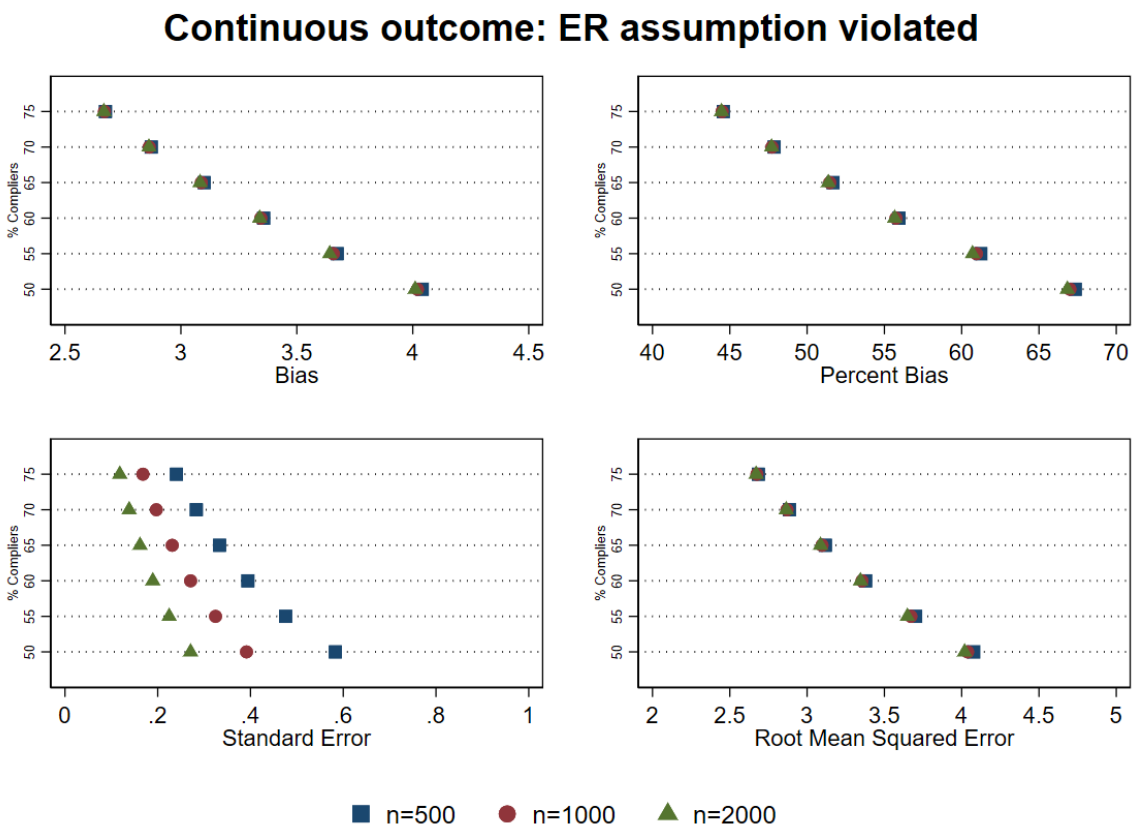

Figure 3.2 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of compliers for the continuous outcome when Assumption I, the exclusion restriction assumption, is violated (scenario B). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

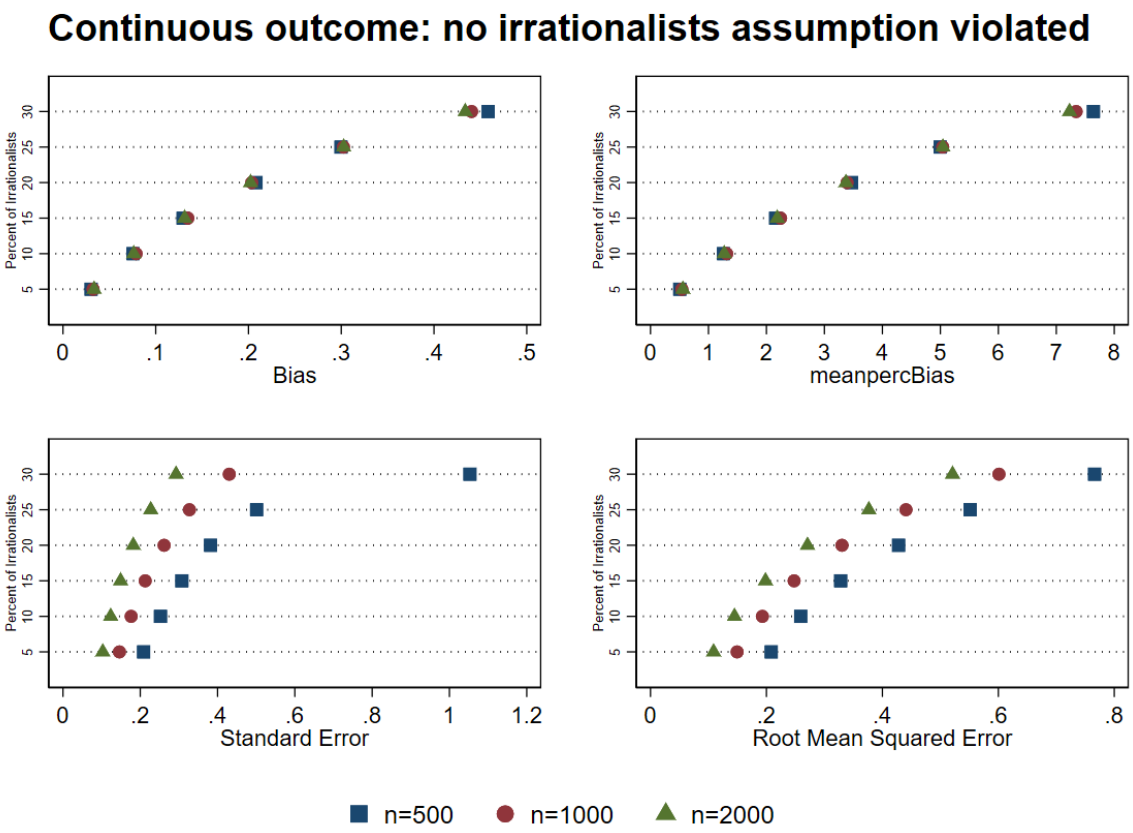

Figure 3.3 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of irrationalists for the continuous outcome when Assumption II.a, the no irrationalists assumption, is violated (scenario C). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

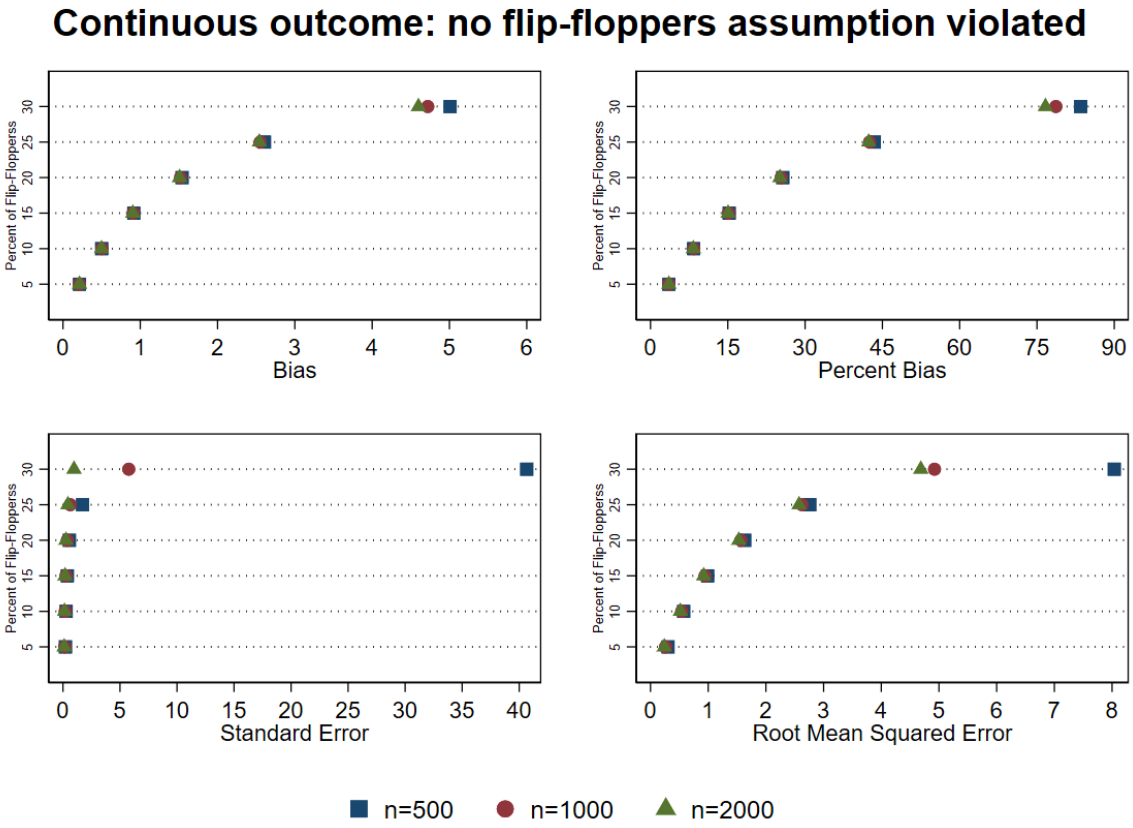

Figure 3.4 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of flip-floppers for the continuous outcome when Assumption II.b, the no flip-floppers assumption, is violated (scenario D). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

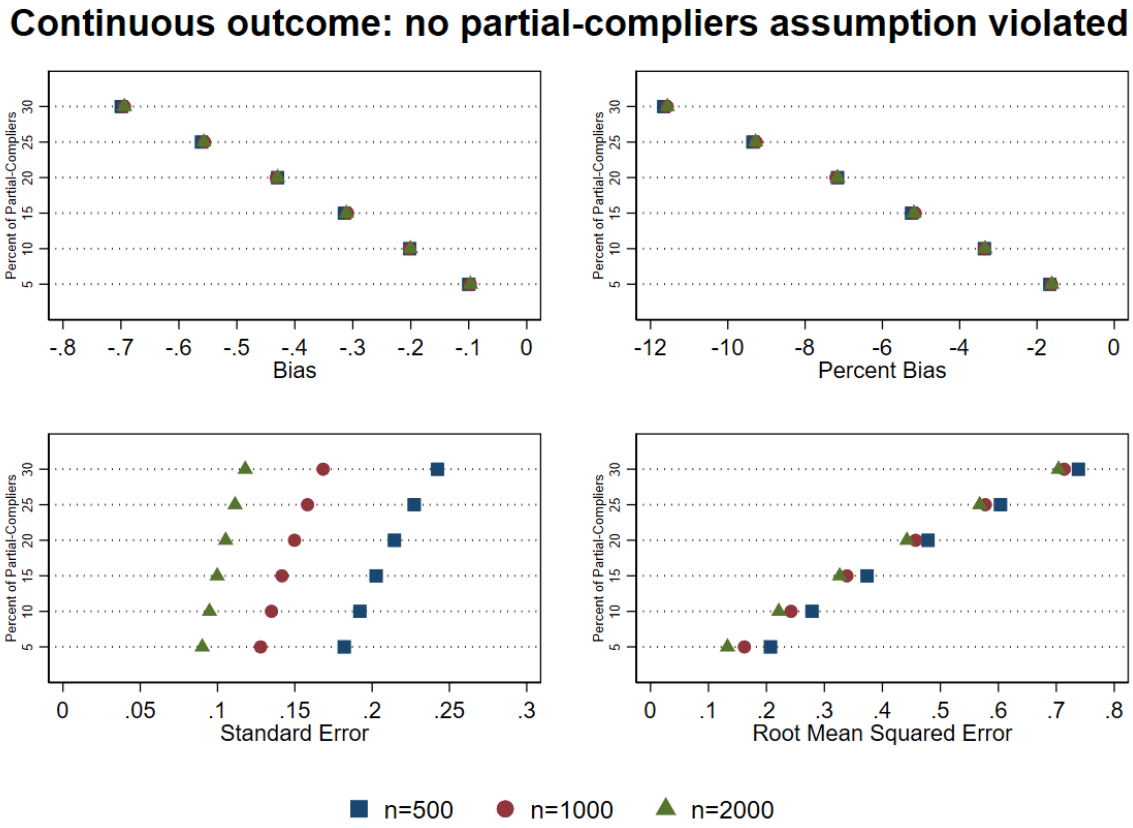

Figure 3.5 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of partial-compliers for the continuous outcome when Assumption III, the no partial-compliers assumption, is violated (scenario E). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

CHAPTER 4

A MULTISITE DESIGN APPROACH

The nonparametric approach identifies our causal estimand of interest by imposing several assumptions. In this chapter, we introduce a method based on the multisite design of a study. Different from the placebo-controlled trials that only has one endogenous variable, our study setup involves two endogenous variables because the treatment received A_i can take three values: 0, 1, or 2. These two endogenous variables are the indicator variables generated from A_i. Namely, one indicates whether individual i takes treatment 1 or not and the other indicates whether takes treatment 2 or not. With one single instrument, i.e., the treatment assignment Z_i and more than one endogenous variables, it is generally not possible to identify any causal effects without further considerations.

However, the identification of causal effects becomes possible in study settings with multiple sites even without looking for extra new instruments. If we assume that the casual structure will be the same across sites for a multisite study, multiple instruments are therefore generated from the site-by-instrument interactions. This idea has the potential to be used in research on causal effects. Raudenbush and Bloom (2015) underlined the importance of learning about the distribution of treatment effects through multisite trials. Reardon and Raudenbush (2013) derived assumptions required to identify the causal effects of multiple mediators using the site-by-treatment instruments. Yuan et al. (2018) proposed a method similar to Reardon and Raudenbush (2013) and estimated the principal strata causal effects rather than causal effects of mediators in a placebo-controlled multisite trial. Our study is an extension to the approach by Yuan et al. (2018) as it involves two active treatments. It also differs from the approach by Reardon and Raudenbush (2013) although there are two mediators (the aforementioned two indicators) in our study. The

effects of the two mediators are from comparisons of either treatment 1 or treatments 2 versus taking nothing. Even if we can take the difference of these two effects to obtain the effect of treatment 2 versus treatment 1, it targets at a population that is not of our interest. We will discuss our method in more detail in the following sections.

4.1 Identification Strategies of the Causal Estimand

We divide the population into the same nine principal strata as described in section 3.1. To identify our causal estimand of interest CACE via the multisite design, we make an additional structural assumption.

Assumption IV (No never-takers): There are no individuals with $(A_i(1) = 0, A_i(2) = 0)$, i.e., $\pi_1 = 0$.

This assumption states that there are no participants who choose to take nothing or to not follow the protocol as required no matter which treatment they are assigned. It can be satisfied if we assume that participants are willing to take the treatment as required when they are eligible to enroll in a trial and sign the consent form.

Under Assumption II and Assumption IV, the nine principal strata are reduced to five, as shown in Table 4.1. Combined with Assumption I (Exclusion Restriction), the overall ITT effect of Z on Y in site k is

$$ITT_{y|k} = ITT_{3|k}\pi_{3|k} + ITT_{4|k}\pi_{4|k} + ITT_{5|k}\pi_{5|k} + ITT_{6|k}\pi_{6|k} + ITT_{9|k}\pi_{9|k}$$
$$= ITT_{3|k}\pi_{3|k} + ITT_{4|k}\pi_{4|k} + CACE_k\pi_{6|k} \quad (4.1)$$

where $\text{ITT}_{y|k} = E\big(Y\big(2, A(2)\big) - Y\big(1, A(1)\big)\big|S = k\big)$ with S representing sites and $k = 1, \ldots, \text{K}$; $\text{ITT}_{g|k} = E\big(Y\big(2, A(2)\big) - Y\big(1, A(1)\big)\big|G = g, S = k\big)$ is the stratum-specific ITT effect of Z on Y in site k and the principal strata g, for $g = 3, 4, 5, 6, 9$; and $\pi_{g|k} = \Pr(G = g|S = k)$ is the site-

specific principal stratum proportion for stratum g in site k. The first equality of equation (4.1) is the decomposition of the overall ITT effect into the weighted stratum-specific ITT effects with weights equal to the principal strata proportions. The last equality is due to Assumption I under which $ITT_{5|k} = ITT_{9|k} = 0$.

Table 4.1 Principal strata classified by A(1) and A(2) for RCTs with two active treatments under Assumption II, the monotonicity assumption, and Assumption IV, the no never-takers assumption.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0			3. Partial-2-complier (P2C) π_3
	A(1)=1	4. Partial-1-complier (P1C) π_4	5. Always-1-taker (A1T) π_5	6. Complier (COMP) π_6
	A(1)=2			9. Always-2-taker (A2T) π_9

As our causal estimand of interest is the CACE for the entire population instead of the CACE for any site, equation (4.1) can be rewritten as

$$\begin{aligned} ITT_{y|k} &= ITT_{3|k}\pi_{3|k} + ITT_{4|k}\pi_{4|k} + CACE_k\pi_{6|k} \\ &= ITT_3\pi_{3|k} + ITT_4\pi_{4|k} + CACE\pi_{6|k} \\ &+ \left(ITT_{3|k} - ITT_3\right)\pi_{3|k} + \left(ITT_{4|k} - ITT_4\right)\pi_{4|k} + (CACE_k - CACE)\pi_{6|k} \\ &= ITT_3\pi_{3|k} + ITT_4\pi_{4|k} + CACE\pi_{6|k} + \epsilon_k \end{aligned} \quad (4.2)$$

where $\epsilon_k = \left(ITT_{3|k} - ITT_3\right)\pi_{3|k} + \left(ITT_{4|k} - ITT_4\right)\pi_{4|k} + (CACE_k - CACE)\pi_{6|k}$; ITT_3, ITT_4 and $CACE$ are the population principal strata causal effects for principal strata 3, 4, and 6 respectively, i.e., $ITT_3 = E(ITT_{3|k})$, $ITT_4 = E(ITT_{4|k})$, and $CACE = E(CACE_k)$.

Equation (4.2) shows a form of multiple linear regression with site-specific ITT effect $ITT_{y|k}$ as the outcome, site-specific principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ as the regressors , and ϵ_k as the error term. As we have K sites, we will have K such equations. We can envision our multisite study setup as follows: suppose we draw K sites from an infinite number of sites, we assume that the site-level parameters $\mathbf{Q} = \left(ITT_{3|k}\ ITT_{4|k}\ CACE_k\ \pi_{3|k}\ \pi_{4|k}\ \pi_{6|k}\right)'$ are independent and identically distributed and follow a distribution with the population-level parameters $\mathbf{M} = (ITT_3\ ITT_4\ CACE\ \pi_3\ \pi_4\ \pi_6)'$ as the mean vector and a 6×6 variance-covariance matrix $\boldsymbol{\Sigma}$ whose element σ_{ij} denotes the covariance between i^{th} and j^{th} element of $\mathbf{Q}$. Given this setup, there is one more assumption needed to identify the CACE using a multiple linear regression. That is, the error term ϵ_k is required to be uncorrelated with the regressors $\pi_{3|k}$, $\pi_{4|k}$, and $\pi_{6|k}$ and has a mean of zero. Specifically, if $cov(\pi_{3|k},\ \epsilon_k) = cov(\pi_{4|k},\ \epsilon_k) = cov(\pi_{6|k},\ \epsilon_k) = 0$ and $\mathrm{E}(\epsilon_k) = 0$, we can identify the population CACE as the coefficient of $\pi_{6|k}$ in the multiple linear regression. A sufficient condition of this requirement yields the following assumption.

Assumption V (Zero correlation): The site-specific principal strata proportions are uncorrelated with the site-specific principal strata causal effects. Specifically,

$$Cov(\pi_{3|k}, ITT_{3|k}) = Cov(\pi_{3|k}, ITT_{4|k}) = Cov(\pi_{3|k}, CACE_k) = 0,$$

$$Cov(\pi_{4|k}, ITT_{3|k}) = Cov(\pi_{4|k}, ITT_{4|k}) = Cov(\pi_{4|k}, CACE_k) = 0,$$

$$Cov(\pi_{6|k}, ITT_{3|k}) = Cov(\pi_{6|k}, ITT_{4|k}) = Cov(\pi_{6|k}, CACE_k) = 0.$$

Assumption V implies that the expected value of the error term ϵ_k is

$$E(\epsilon_k) = E\left(\left(ITT_{3|k} - ITT_3\right)\pi_{3|k} + \left(ITT_{4|k} - ITT_4\right)\pi_{4|k} + (CACE_k - CACE)\pi_{6|k}\right)$$

$$= E\left(ITT_{3|k}\pi_{3|k}\right) + E\left(ITT_{4|k}\pi_{4|k}\right) + E\left(CACE_k\pi_{6|k}\right)$$

$$- ITT_3 E(\pi_{3|k}) - ITT_4 E(\pi_{4|k}) - E(\pi_{6|k}) CACE$$

$$= cov\left(\pi_{3|k}, ITT_{3|k}\right) + E(\pi_{3|k}) E\left(ITT_{3|k}\right)$$

$$+ cov\left(\pi_{4|k}, ITT_{4|k}\right) + E(\pi_{4|k}) E\left(ITT_{4|k}\right)$$

$$+ cov\left(\pi_{6|k}, CACE_k\right) + E(\pi_{6|k}) E(CACE_k)$$

$$- ITT_3 E(\pi_{3|k}) - ITT_4 E(\pi_{4|k}) - E\left(\pi_{6|k}\right) CACE$$

$$= 0$$

Likewise, Assumption V also implies that $cov\left(\pi_{3|k},\ \epsilon_k\right) = cov\left(\pi_{4|k},\ \epsilon_k\right) = cov\left(\pi_{6|k},\ \epsilon_k\right) = 0$. We provide the proof for $cov\left(\pi_{3|k},\ \epsilon_k\right) = 0$ below. The same logic applies to the other two conditions $cov\left(\pi_{4|k},\ \epsilon_k\right) = 0$ and $cov\left(\pi_{6|k},\ \epsilon_k\right) = 0$.

$$cov\left(\pi_{3|k}, \epsilon_k\right) = cov(\pi_{3|k}, \left(ITT_{3|k} - ITT_3\right)\pi_{3|k} + \left(ITT_{4|k} - ITT_4\right)\pi_{4|k}$$

$$+ (CACE_k - CACE)\pi_{6|k})$$

$$= cov\left(\pi_{3|k}, ITT_{3|k}\pi_{3|k}\right) + cov\left(\pi_{3|k}, ITT_{4|k}\pi_{4|k}\right) + cov\left(\pi_{3|k}, CACE_k\pi_{6|k}\right)$$

$$- ITT_3 cov\left(\pi_{3|k}, \pi_{3|k}\right) - ITT_4 cov\left(\pi_{3|k}, \pi_{4|k}\right) - CACE cov\left(\pi_{3|k}, \pi_{6|k}\right)$$

$$= E\left(\pi_{3|k}^2 ITT_{3|k}\right) - E\left(\pi_{3|k}\right) E\left(ITT_{3|k}\pi_{3|k}\right)$$

$$+ E\left(\pi_{3|k} ITT_{4|k}\pi_{4|k}\right) - E\left(\pi_{3|k}\right) E\left(ITT_{4|k}\pi_{4|k}\right)$$

$$+ E\left(\pi_{3|k} CACE_k\pi_{6|k}\right) - E\left(\pi_{3|k}\right) E\left(CACE_k\pi_{6|k}\right)$$

$$- ITT_3 var\left(\pi_{3|k}\right) - ITT_4 cov\left(\pi_{3|k}, \pi_{4|k}\right) - CACE cov\left(\pi_{3|k}, \pi_{6|k}\right)$$

$$= E\left(\pi_{3|k}^2\right) E(ITT_{3|k}) - E\left(\pi_{3|k}\right)\left(cov\left(\pi_{3|k}, ITT_{3|k}\right) + E\left(ITT_{3|k}\right) E\left(\pi_{3|k}\right)\right)$$

$$+ E\left(\pi_{3|k}\pi_{4|k}\right) E(ITT_{4|k}) - E\left(\pi_{3|k}\right)\left(cov\left(\pi_{4|k}, ITT_{4|k}\right) + E\left(ITT_{4|k}\right) E\left(\pi_{4|k}\right)\right)$$

$$+ E\left(\pi_{3|k}\pi_{6|k}\right) E(CACE_k) - E\left(\pi_{3|k}\right)\left(cov\left(\pi_{6|k}, CACE_k\right) + E(CACE_k) E\left(\pi_{6|k}\right)\right)$$

$$- ITT_3 var\left(\pi_{3|k}\right) - ITT_4 cov\left(\pi_{3|k}, \pi_{4|k}\right) - CACE cov\left(\pi_{3|k}, \pi_{6|k}\right)$$

$$= E(ITT_{3|k})\left(E(\pi_{3|k}^2) - E(\pi_{3|k})^2\right)$$

$$+ E(ITT_{4|k})\left(E(\pi_{3|k}\pi_{4|k}) - E(\pi_{3|k})E(\pi_{4|k})\right)$$

$$+ E(CACE_k)(E(\pi_{3|k}\pi_{6|k}) - E(\pi_{3|k})E(\pi_{6|k}))$$

$$- ITT_3 var(\pi_{3|k}) - ITT_4 cov(\pi_{3|k}, \pi_{4|k}) - CACE cov(\pi_{3|k}, \pi_{6|k})$$

$$= 0$$

Assumption V is a nontrivial assumption and may not be empirically verifiable. We will explore the sensitivity of this assumption through simulation studies in section 4.4. Combining all the assumptions mentioned above results in Proposition II.

Proposition II: For a RCT with two active treatments subject to nonadherence, the complier average causal effect (CACE) can be identified under Assumption I, II, IV, and V via a multisite design.

4.2 Estimation of the Causal Estimand

To identify CACE using equation (4.2), it requires us to first identify the proportions of individuals in principal strata 3, 4, and 6 for each site, i.e., $\pi_{3|k}$, $\pi_{4|k}$, and $\pi_{6|k}$. In other words, we need to identify which principal stratum each individual belongs to in site k. As the principal strata are determined by the potential treatment received $A(Z)$ and it is impossible to observe both $A(1)$ and $A(2)$ for one individual simultaneously, thus it is impossible to determine to which principal stratum an individual belongs. We can, however, estimate these proportions by using the observed quantities. As the observed treatment assignment Z takes two values and the observed treatment receipt A takes three values, there are in total six possible Z and A combinations that we can observe. These six combinations are mixtures of the five principal strata shown in Table 4.1.

We can therefore estimate the principal strata proportions through the proportions of the observed combinations.

Let $\mathrm{p_{a|zk}} = \Pr(A = a|Z = z, S = k)$ be the probabilities of the six observed combinations in site k where $\mathrm{a} = 0,1,2$; $\mathrm{z} = 1,2$; and $\sum_{a=0,1,2} p_{a|zk} = 1$. We can calculate $\mathrm{p_{a|zk}}$ by simply taking the ratio of $\mathrm{N_{a|zk}}$ and $\mathrm{N_{z|k}}$ where $\mathrm{N_{a|zk}}$ is the number of individuals that are assigned to treatment z and take treatment a in site k; $\mathrm{N_{z|k}}$ is the total number of individuals that are assigned to treatment z in site k. Then, we can immediately estimate $\mathrm{\pi_{3|k}}$, $\mathrm{\pi_{4|k}}$, $\mathrm{\pi_{5|k}}$ and $\mathrm{\pi_{9|k}}$ from $\mathrm{p_{0|1k}}$, $\mathrm{p_{0|2k}}$, $\mathrm{p_{1|2k}}$ and $\mathrm{p_{2|1k}}$ as follows

$$\hat{\pi}_{3|k} = \mathrm{p_{0|1k}} = \frac{\mathrm{N_{0|1k}}}{\mathrm{N_{1|k}}},$$

$$\hat{\pi}_{4|k} = \mathrm{p_{0|2k}} = \frac{\mathrm{N_{0|2k}}}{\mathrm{N_{2|k}}},$$

$$\hat{\pi}_{5|k} = \mathrm{p_{1|2k}} = \frac{\mathrm{N_{1|2k}}}{\mathrm{N_{2|k}}},$$

$$\hat{\pi}_{9|k} = \mathrm{p_{2|1k}} = \frac{\mathrm{N_{2|1k}}}{\mathrm{N_{1|k}}},$$

because randomization ensures that the principal strata proportions are the same across treatment arms. Take estimating the proportion of partial-2-compliers ($\mathrm{\pi_{3|k}}$) as an example. First, under Assumptions II and IV, we believe that those who are assigned to treatment 1 and take nothing are all partial-2-compliers. Then, we obtain the proportion of partial-2-compliers from those who are assigned to treatment 1 for site k, which is $\mathrm{p_{0|1k}}$. Due to randomization, the proportion of partial-2-compliers in the entire population will be equal to the proportion of partial-2-compliers in each treatment assignment arm of the site, i.e., $\Pr(A_i(1) = 0, A_i(2) = 2|S = k) = \Pr(A_i(1) = 0, A_i(2) = 2|S = k, Z_i = z)$ for $z = 1, 2$. Therefore, we can estimate $\mathrm{\pi_{3|k}}$ from $\mathrm{p_{0|1k}}$ for site k. Similarly, to estimate the proportion of compliers, we first observe those who are assigned to

treatment 2 and also take treatment 2 and calculate this quantity as $\mathrm{p}_{2|2\mathrm{k}} = \frac{\mathrm{N}_{2|2\mathrm{k}}}{\mathrm{N}_{2|\mathrm{k}}}$. These individuals are a mixture of partial-2-compliers (principal stratum 3), compliers (principal stratum 6) and always-2-takers (principal stratum 9). As we have obtained $\hat{\pi}_{3|k}$ and $\hat{\pi}_{9|k}$, the proportion of compliers is estimated by

$$\hat{\pi}_{6|k} = \mathrm{p}_{2|2\mathrm{k}} - \hat{\pi}_{3|k} - \hat{\pi}_{9|k} = \mathrm{p}_{2|2\mathrm{k}} - \mathrm{p}_{0|1\mathrm{k}} - \mathrm{p}_{2|1\mathrm{k}} = \frac{\mathrm{N}_{2|2\mathrm{k}}}{\mathrm{N}_{2|\mathrm{k}}} - \frac{\mathrm{N}_{0|1\mathrm{k}}}{\mathrm{N}_{1|\mathrm{k}}} - \frac{\mathrm{N}_{2|1\mathrm{k}}}{\mathrm{N}_{1|\mathrm{k}}}$$

This formula also provides a way of falsifying Assumption IV. Specifically, if we use the above formula to derive the estimated population principal stratum proportion $\hat{\pi}_6$, there is a chance that $\hat{\pi}_6$ can be negative, therefore enabling us to falsify Assumption IV if needed.

After identifying the principal strata proportions, we also need to identify the ITT effect of Z on Y for each site k. We know that $ITT_{y|k} = E\big(Y\big(2, A(2)\big) - Y\big(1, A(1)\big)\big|S = k\big) = E(Y|Z = 2, S = k) - E(Y|Z = 1, S = k)$ where the second equality is due to randomization and consistency. Then, we can use the sample average to estimate $E(Y|Z = z, S = k), z = 1,2$:

$$\widehat{ITT}_{y|k} = \frac{\sum_{i\epsilon\{i:Z_i=2\}} Y_{i|k}}{N_{2|k}} - \frac{\sum_{i\epsilon\{i:Z_i=1\}} Y_{i|k}}{N_{1|k}}$$

where $Y_{i|k}$ is the observed outcome for individual i in site k. With the estimated ITT effects and the estimated principal strata proportions, we can use equation (4.3) to identify our causal estimand of interest via a multiple linear regression, that is,

$$\widehat{ITT}_{y|k} = ITT_3\hat{\pi}_{3|k} + ITT_4\hat{\pi}_{4|k} + CACE\hat{\pi}_{6|k} + \hat{\epsilon}_k \qquad (4.3)$$

4.3 The Problems of Omitted Variables and Measurement Errors

As our multisite design approach relies on Assumption V, the zero correlation assumption, violation of this assumption will result in bias in the estimates, which is the so-called omitted-

variable bias. In addition, because we use the estimated site-specific principal strata proportions and the estimated site-specific ITT effects to identify CACE rather than the true principal strata proportions and the true ITT effect, the measurement errors in these quantities will also bias the estimates. For exposition, we explore the issues of measurement errors and omitted variables in matrix format in this section.

Generally, consider the true model

$$\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{U\gamma} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y} = (Y_1 \quad \cdots \quad Y_n)'$ is a $n \times 1$ column vector of the true outcomes; $\boldsymbol{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}$ is a $n \times p$ matrix of the regressors; $\boldsymbol{U} = \begin{pmatrix} U_{11} & \dots & U_{1q} \\ \vdots & \ddots & \vdots \\ U_{n1} & \dots & U_{nq} \end{pmatrix}$ is a $n \times q$ matrix of the unobserved confounding variables that are correlated with some or all components of $\boldsymbol{X}$; $\boldsymbol{\theta} = (\theta_1 \quad \cdots \quad \theta_p)'$ is a $p \times 1$ column vector for the coefficients of $\boldsymbol{X}$; $\boldsymbol{\gamma} = (\gamma_1 \quad \cdots \quad \gamma_q)'$ is a $q \times 1$ column vector for the coefficients of $\boldsymbol{U}$; and $\boldsymbol{\epsilon} = (\epsilon_1 \quad \cdots \quad \epsilon_n)'$ is a $n \times 1$ column vector of the error terms that follow a standard normal distribution and are independent of $\boldsymbol{X}$ and $\boldsymbol{U}$. Now, let $\boldsymbol{V}$ be the observed outcome vector that measures $\boldsymbol{Y}$ with classic measurement error $\boldsymbol{\delta}$, i.e., $\boldsymbol{V} = \boldsymbol{Y} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1 \quad \cdots \quad \delta_n)'$ is a $n \times 1$ column vector representing the measurement errors in $\boldsymbol{V}$ and assumed to be independent of $\boldsymbol{Y}$. Similarly, let $\boldsymbol{W}$ be the observed regressors matrix that measures $\boldsymbol{X}$ with classic measurement errors $\boldsymbol{e}$, i.e., $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{e}$, where $\boldsymbol{e} = \begin{pmatrix} e_{11} & \dots & e_{1p} \\ \vdots & \ddots & \vdots \\ e_{n1} & \dots & e_{np} \end{pmatrix}$ is a $n \times p$ matrix representing the measurement errors in $\boldsymbol{W}$ and assumed to be independent of $\boldsymbol{X}$. Due to the unobserved confounding variables (the omitted variables) and the measurement errors in both outcomes and regressors, the actual regression model becomes

$$\boldsymbol{V} = \boldsymbol{W\theta}^* + \boldsymbol{\epsilon}^*$$

Thus, the ordinary least square (OLS) estimator of $\boldsymbol{\theta}^*$ is

$$\begin{aligned}
\widehat{\boldsymbol{\theta}^*} &= (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{V} \\
&= (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{Y} + \boldsymbol{\delta}) \\
&= (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{X\theta} + \boldsymbol{U\gamma} + \boldsymbol{\epsilon} + \boldsymbol{\delta}) \\
&= (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X\theta} + (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{U\gamma} + (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{\epsilon} + \boldsymbol{\delta}) \qquad (4.4)
\end{aligned}$$

Taking expectation of equation (4.4), we have

$$\begin{aligned}
E\left(\widehat{\boldsymbol{\theta}^*}\right) &= E\left(E\left(\widehat{\boldsymbol{\theta}^*}|\boldsymbol{W}\right)\right) \\
&= E(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X\theta} + (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{U\gamma} + (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{\epsilon} + \boldsymbol{\delta})|\boldsymbol{W})) \\
&= E\left(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{W} - \boldsymbol{e})\boldsymbol{\theta}|\boldsymbol{W})\right) + E\left(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{U\gamma}|\boldsymbol{W})\right) \\
&+ E\left(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'(\boldsymbol{\epsilon} + \boldsymbol{\delta})|\boldsymbol{W})\right) \\
&= \boldsymbol{\theta} - E\left(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{e\theta}|\boldsymbol{W})\right) + E\left(E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{U\gamma}|\boldsymbol{W})\right) \\
&= \boldsymbol{\theta} + E((\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{U\gamma} - (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{e\theta}) \\
&= \boldsymbol{\theta} + E((\boldsymbol{W}'\boldsymbol{W})^{-1}(\boldsymbol{W}'\boldsymbol{U\gamma} - \boldsymbol{W}'\boldsymbol{e\theta})) \\
&= \boldsymbol{\theta} + E((\boldsymbol{W}'\boldsymbol{W})^{-1}((\boldsymbol{X} + \boldsymbol{e})'\boldsymbol{U\gamma} - (\boldsymbol{X} + \boldsymbol{e})'\boldsymbol{e\theta})) \\
&= \boldsymbol{\theta} + E((\boldsymbol{W}'\boldsymbol{W})^{-1}(\boldsymbol{X}'\boldsymbol{U\gamma} - \boldsymbol{e}'\boldsymbol{e\theta})) \qquad (4.5)
\end{aligned}$$

The last item of the third equality equals to zero because the conditional expectations of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are zero. The last equality holds because $\boldsymbol{e}$ is independent of $\boldsymbol{U}$ and $\boldsymbol{X}$. Thus, we can derive the bias formula from equation (4.5) when there is omitted variables as well as measurement errors in regressors and outcomes. That is

$$\text{Bias} = E\left(\widehat{\boldsymbol{\theta}^*}\right) - \boldsymbol{\theta} = E((\boldsymbol{W}'\boldsymbol{W})^{-1}(\boldsymbol{X}'\boldsymbol{U\gamma} - \boldsymbol{e}'\boldsymbol{e\theta})) \qquad (4.6)$$

To apply the general bias formula in equation (4.6) into our study setting, first we transform the notations in section 4.1 into matrix format. Let $\boldsymbol{\pi} = \begin{pmatrix} \pi_{3|1} & \pi_{4|1} & \pi_{6|1} \\ \vdots & \vdots & \vdots \\ \pi_{3|K} & \pi_{4|K} & \pi_{6|K} \end{pmatrix}$ be the $K \times 3$ matrix representing the true principal strata proportions and the population principal strata causal effects vector $\boldsymbol{\beta} = (ITT_3 \quad ITT_4 \quad CACE)'$ is the coefficients of $\boldsymbol{\pi}$. Correspondingly, $\hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_{3|1} & \hat{\pi}_{4|1} & \hat{\pi}_{6|1} \\ \vdots & \vdots & \vdots \\ \hat{\pi}_{3|K} & \hat{\pi}_{4|K} & \hat{\pi}_{6|K} \end{pmatrix}$ will be the $K \times 3$ matrix for the estimated principal strata proportions. Thus, the measurement error matrix $\boldsymbol{e}$ becomes $\boldsymbol{e} = \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}$. We still use the notation $\boldsymbol{U}$ for the unobserved site-level covariates with $\boldsymbol{\gamma}$ as its coefficient vector, and the dimensions of $\boldsymbol{U}$ is $K \times 1$. With these notations, plugging $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{W}$ and $\boldsymbol{\pi}$ for $\boldsymbol{X}$ into equation (4.6), we have

$$\text{Bias} = \text{E}((\hat{\boldsymbol{\pi}}'\hat{\boldsymbol{\pi}})^{-1}(\boldsymbol{\pi}'\mathbf{U}\boldsymbol{\gamma} - \boldsymbol{e}'\boldsymbol{e}\boldsymbol{\theta})) \tag{4.7}$$

for our multisite study. This formula implies that the bias will not only depend on the variance of $\hat{\boldsymbol{\pi}}$ which is the sum of the variance of $\boldsymbol{\pi}$ and the variance of $\boldsymbol{e}$ but also on the covariance between $\boldsymbol{\pi}$ and $\boldsymbol{U}$. Specifically, the bias will be impacted by the variance of the true principal strata proportions, the variance of the measurement errors in estimating the true principal strata proportions, as well as the covariance between the true principal strata proportions and the unobserved site-level covariates. These implications will be explored through our simulations.

4.4 Simulation Studies

4.4.1 Simulation Setup

We design a two-stage data generating process for our simulations. For the site-level data, we use a truncated multivariate normal distribution with mean vector $\boldsymbol{m} = (0.05 \quad 0.05 \quad 0.5)'$ and

variance-covariance matrix $\boldsymbol{\Gamma} = \begin{pmatrix} 0.01 & & \\ -0.0025 & 0.01 & \\ -0.0025 & -0.0025 & 0.01 \end{pmatrix}$ to generate the site-specific principal strata proportions of partial-2-compliers, partial-1-compliers and compliers, i.e., $\pi_{3|k}, \pi_{4|k}$, and $\pi_{6|k}$ for K sites. We truncate the three principal proportions with a lower limit $\boldsymbol{L} = (0.01 \quad 0.01 \quad 0.1)'$ and a upper limit $\boldsymbol{U} = (0.15 \quad 0.15 \quad 0.7)'$. With the generated $\pi_{3|k}, \pi_{4|k}$, and $\pi_{6|k}$, we then draw $\pi_{5|k}$ for always-1-takers from a uniform distribution $(0, 1 - \pi_{3|k} - \pi_{4|k} - \pi_{6|k})$ and calculate $\pi_{9|k}$ for always-2-takers from $1 - \pi_{3|k} - \pi_{4|k} - \pi_{6|k} - \pi_{5|k}$. For the number of sites K, we consider three different values for our simulations, that is, $K = 50$, 100, or 200.

Given the site-level quantities, we next generate the individual-level data for each site. First, we need to determine the site size of each site, that is, the number of individuals in a site. Although site sizes usually vary by sites in practice, we explore both situations where the site size is fixed or varies across sites. We consider the fixed site size as a simple and special case for the real world and look at three values: 25, 50, or 100 individuals for every site. For varied site size, we use a Poisson distribution with $\lambda = 10$ to generate the site size for each site and scale it by 5. Then, with the known site size, we generate the principal stratum G_i for each individual i. Specifically, we first generated a variable T from a uniform distribution $U(0, 1)$. Then, we use T to inverse transform a multinomial distribution with probabilities $\pi_{g|k}, g = 3,4,5,6,9$ for the principal stratum G_i. We use the same variable T to inverse transform a binomial distribution for an individual-level covariate U_i which will therefore be correlated with the principal stratum G_i. We aggregate U_i to obtain a site-level covariate U_s for later use.

For the treatment assignment status Z_i, we use a Bernoulli distribution with a probability of 0.5 to generate the two active randomization arms. Thus, given G_i and Z_i, we can obtain the

observed treatment received A_i in the same way as we did in Step 3) of the nonparametric simulation setup.

To better represent the real world, we generate our potential outcomes $\left(Y_i(1, A_i(1)), Y(2, A_i(2))\right)$ for each individual through random effect models. Specifically, for the continuous outcomes, we use the formula

$$Y_i(1, A_i(1)) = \alpha_3 g_{3i} + \alpha_4 g_{4i} + \alpha_6 g_{6i} + \lambda U_s + \epsilon_b + \epsilon_{wi}$$

$$Y_i(2, A_i(2)) = Y_i(1, A_i(1)) + \beta_3 g_{3i} + \beta_4 g_{4i} + \beta_6 g_{6i} + \gamma U_s$$

where g_{ji} is an indicator with $g_{ji} = 1$ if individual i is in principal stratum j and 0 otherwise. ϵ_b is generated at the site level and represents the between-site variation whereas ϵ_{wi} is an individual-level random error which is also known as the within-site variation. Both ϵ_b and ϵ_{wi} follow a standard normal distribution $N(0,1)$. U_s is the site-level covariate that we generated previously. Now, it shows clearly that U_s is a confounding variable that is associated with both the site-level principal strata proportions and the potential outcomes. Similarly, we generate the binary potential outcomes from a logistic regression:

$$logit\left(\Pr\left(Y_i(1, A(1)) = 1\right)\right) = \alpha_3 g_3 + \alpha_4 g_4 + \alpha_6 g_6 + \lambda U_s + \epsilon_b$$

$$logit\left(\Pr\left(Y_i(2, A(2)) = 1\right)\right) = logit\left(\Pr\left(Y_i(1, A(1)) = 1\right)\right) + \beta_3 g_3 + \beta_4 g_4 + \beta_6 g_6 + \gamma U_s$$

With the treatment assignment status Z_i and the potential outcomes generated above, the observed outcome Y_i will be equal to $Y_i(z, A_i(z))$ if z=1 or 2.

After all data are generated, we can use the multiple linear regression to obtain the estimate of our causal estimand of interest. As discussed in section 4.3, our estimator may be biased when measurement errors and omitted variables exist. We investigate these problems by considering three types of estimators: oracle, naïve and bootstrap. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression in equation (4.2). The

naïve estimator is obtained by fitting the regression in equation (4.3) using the estimated principal strata proportions. We also include a bootstrap estimator whose estimate is the average of the naïve estimates from resampling the site-level quantities. We use the estimated ITT effects $\widehat{ITT}_{y|k}$ as the outcome of the multiple linear regression for all three estimators for consistency as measurement errors in regressors rather than outcomes are our major concern. For varied site size, we fit both the unweighted and weighted regression. Three types of weights are considered: site size, within-site variation, and complier size per site.

Table 4.2 Summary of the parameter values used to generate the potential outcomes for the multisite design approach.

Analysis	Outcome Type	Parameters							
		α_3	α_4	α_6	β_3	β_4	β_6	λ	γ
Performance Analysis	Continuous	-3	-4	-2	8	-7	6	0	0
	Binary	-3	-1	-1	4	-3	1.51	0	0
Sensitivity Analysis	Continuous	-3	-4	-2	8	-7	6	2	1 or -1
	Binary	-3	-1	-1	4	-3	1.51	2	1 or -1

We conduct two analyses for both the continuous outcomes and the binary outcomes. One is the performance analysis that evaluates the performance of our estimators when all assumptions required for the identification of the causal estimand are met. The other one is the sensitivity analysis when Assumption V is violated. We set the true values of CACE to be 6 for the continuous case and 0.3 for the binary case. Parameter values that are used for simulating the data are the same for the fixed site size and the varied site size. We summarize these values in Table 4.2. For the performance analysis, we set $\lambda = \gamma = 0$ because in this way no confounding variable U_s exists

and Assumption V is therefore satisfied. For sensitivity analysis, we fix λ to be 2 but vary the sign of γ as γ plays a role in the bias formula as in equation (4.7). For each scenario, we repeat our simulation 10000 times.

4.4.2 Results

We discuss the performance and sensitivity of our estimator using the same performance metrics that are described in section 3.4.2. We only display the results for the continuous outcomes here. Results for the binary outcomes are similar (see Appendix B). For fixed site size, we consider nine scenarios based on the combinations of number of sites and site sizes. For varied site size, there are twelve scenarios based on the combinations of types of weights (including unweighted) and number of sites.

4.4.2.1 Performance Analysis

Figure 4.1 includes the bias, percent bias, mean standard errors, root mean squared errors, ratios of mean standard errors and empirical standard errors, and coverage rates for samples with fixed site size. In each panel, the dotted lines represent the nine scenarios. For example, label "s=50, n=25" indicates that the scenario has 50 sites and 25 individuals in each site. The top two panels of Figure 4.1 show that the oracle estimator is unbiased across all scenarios. This is as expected as the OLS estimator is unbiased when there are no omitted variables as well as no measurement errors in the regressors. The naïve estimator and the bootstrap estimator are slightly biased due to the measurement errors, which can be explained by equation (4.7). As there is no omitted variable issue in the performance analysis, the term $\boldsymbol{\pi}'\mathbf{U}\boldsymbol{\gamma}$ will not exist. Then, the bias will only be associated with the variance of the true principal strata proportions and the variance of the

measurement errors. Specifically, the bias will decrease as the variance of the measurement errors decreases whereas it will increase as the variance of the regressors decreases. This pattern is reflected in our results. In Figure 4.1, we observe that the bias becomes smaller when we increase the site size given a fixed number of sites. This is because the estimated principal strata proportions will be closer to the true proportions when the site size becomes larger. As a result, the variance of the measurement errors will become smaller. We expect an increase in the bias if the number of sites increases while the site size is controlled. However, it seems that this pattern is not significant especially for larger site sizes. For example, the percent bias looks almost the same when the site size is 100, ranging from -0.86% when $s = 50$ to -0.89% when $s = 200$ for the naïve estimator. This is reasonable because 50 is a relatively big enough sample size for the linear regression, further increasing the sample size will not make a significant difference in reducing the dispersion of the true principal strata proportions.

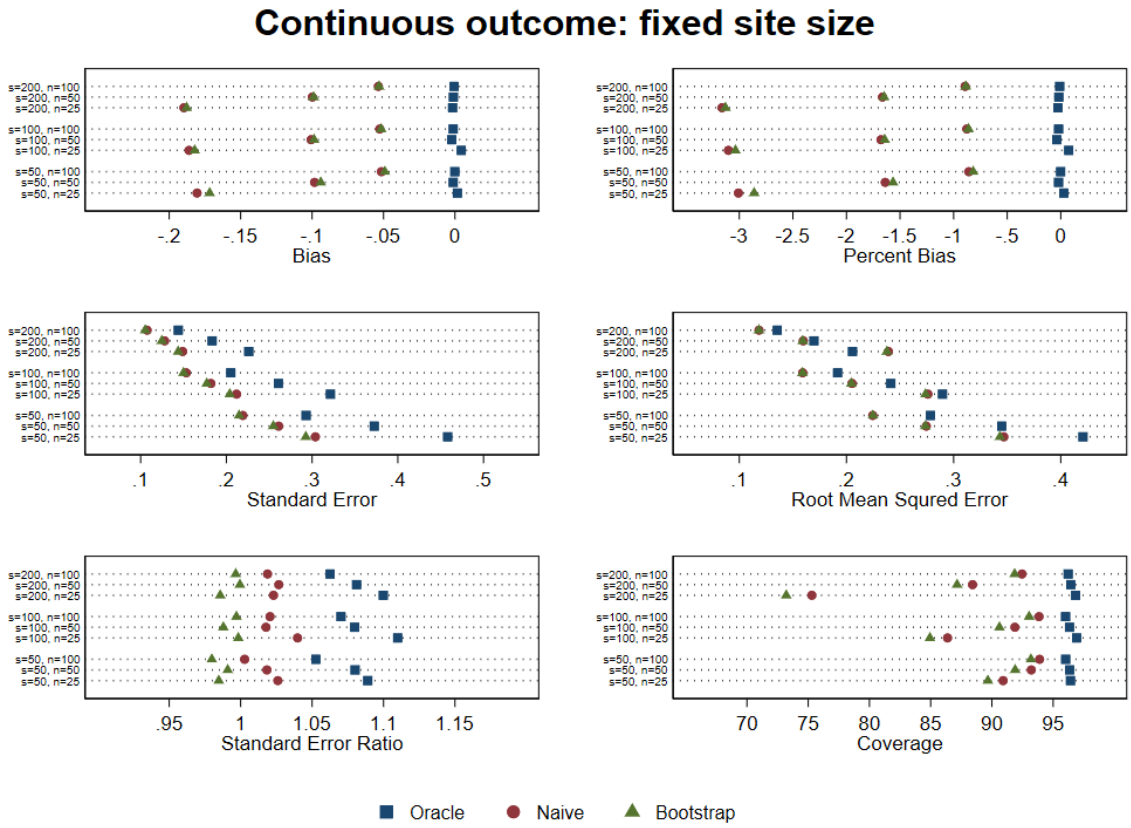

Figure 4.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the nine scenarios based on number of sites s and site sizes n for the continuous outcome when site size is fixed. For example, label "s=50, n=25" indicates that the scenario has 50 sites and 25 individuals in each site. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities.

A clear pattern can be observed for the mean standard errors of the three estimators. Namely, the estimators become more precise as the number of sites and the site size increases. One interesting finding is that the mean standard errors of the oracle estimator is always bigger than that of the naïve or the bootstrap estimator when the number of sites and the site size is fixed. This is plausible because the variance of the OLS estimator will be inversely related to the variance of the regressors holding everything else constant. As the variability in regressors is minimal when estimating the oracle estimator compared to the other two estimators, the mean standard errors of the oracle estimator will as a result be the largest among the three estimators within the same scenario. The root mean squared errors have a similar pattern as the mean standard errors across scenarios. The more sites, the bigger the site size, the smaller the root mean squared errors of the estimators.

The mean standard errors of the naïve and the bootstrap estimators are close to the standard deviation of the estimated CACE as the standard error ratios for these two estimators lie around one as shown at the lower left corner of Figure 4.1. On the other hand, the mean standard errors of the oracle estimator slightly overestimate the empirical standard errors. This suggests that the mean standard errors of the oracle estimator may be a little conservative, which is consistent with the implications found for the mean standard errors. The lower right corner of Figure 4.1 shows the coverage rates for the three estimators. The oracle estimator has a higher coverage rate that is around 96% across the nine scenarios than the other two estimators. Neither the naïve estimator nor the bootstrap estimator can provide valid coverage rates for the effect of interest. The coverage rates of these two estimators move away from 95% as the number of sites increases given the site size, and move towards 95% as the site size increases given the number of sites.

Similar findings for the varied site size are shown in Figure 4.2. As in Figure 4.1, the dotted lines represent the twelve scenarios for the varied site size with UN standing for unweighted, SS for weighting by the site size, WV for weighting by the within-site variation and COMP for weighting by the number of compliers. For every type of weights, the oracle estimator is unbiased as expected. There is a slight and insignificant increase in the bias of the naïve estimator and the bootstrap estimator when there are more sites within each type of weights. Weighting by the within-site variation gives the largest bias among the four types of weights. This is explainable because weighting in this way compromises the variability of the estimated principal strata proportions and therefore results in larger bias.

The mean standard errors and the root mean squared errors of the three estimators for the varied site size show the same patterns as for the fixed site size. The mean standard errors of the naïve estimator and the bootstrap estimator are always similar to each other within each scenario for the same type of weights except when using the within-site variation as weights. In that case, the mean standard errors of the naïve estimator are smaller than those of the bootstrap estimator. This is also reflected in the standard error ratio panel of Figure 4.2. The smaller mean standard errors of the naïve estimator when weighting by the within-site variation results in a severe underestimate for the empirical standard errors, leading to coverage rates that are much lower than 95%. In general, the coverage rates of the oracle estimator are around 95%, however, the coverage rates of the naïve estimator and the bootstrap estimator are all below 95% as they are biased due to measurement errors.

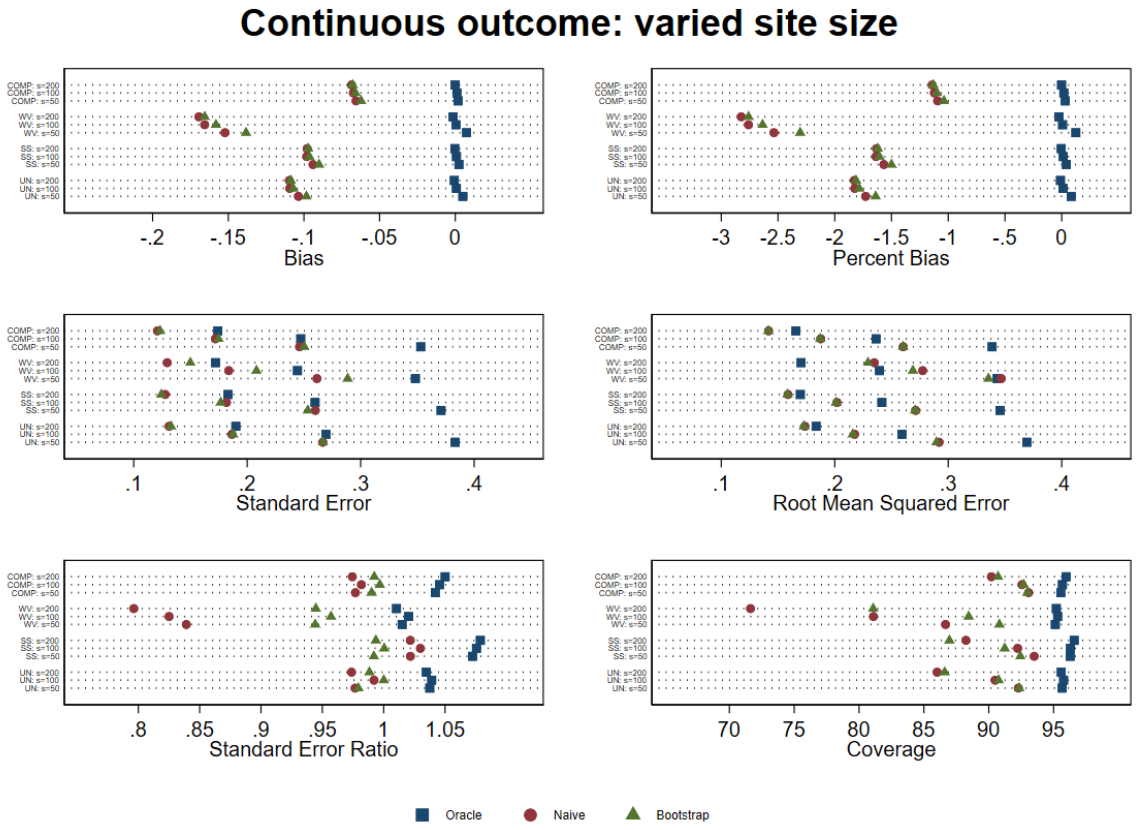

Figure 4.2 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites s for the continuous outcome when site size is varied. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities.

4.4.2.2 Sensitivity Analysis

When there are omitted variables that are associated with the site-specific principal strata proportions and the site-specific principal strata ITT effects, Assumption V will be violated. For our simulations, we only include one omitted variable U_s for simplicity. As shown in Equation (4.7), the bias will be affected by the product of the covariance between the omitted variable U_s and the principal strata proportions $\boldsymbol{\pi}$ and the coefficient of the omitted variable γ. Therefore, as our simulations have negative covariance between U_s and $\boldsymbol{\pi}$ for all scenarios, we consider two values for γ with opposite signs, i.e., $\gamma = 1$ or $\gamma = -1$, to see how our estimators perform when there is deviation from the zero correlation assumption. We compare three cases in Table 4.3 for the fixed site size and Table 4.4 for the varied site size with C1 referring to the case when Assumption V holds ($\lambda = \gamma = 0$, reference case), C2 referring to the case when $\lambda = 2$ and $\gamma = 1$, and C3 referring to the case when $\lambda = 2$ and $\gamma = -1$. The results for the oracle estimator show the influence directly from violation of Assumption V. However, the results for the naïve estimator and the bootstrap estimator are more practical because they combine the impact of the two most frequently occurred situations in the practical setting, i.e., the influence from the omitted variable and the measurement errors in estimating the principal strata proportions.

Table 4.3 Sensitivity of the multisite design estimator across the nine scenarios based on number of sites and site sizes for the continuous outcome when site size is fixed if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$.

Measure	# of Sites	Site Size	Oracle			Naïve			Bootstrap		
			C1	C2	C3	C1	C2	C3	C1	C2	C3
Bias ($\times 100$)	50	25	0.17	-0.69	0.70	-0.18	-3.06	-1.82	-0.17	-3.05	-1.82
		50	-0.12	-0.63	0.62	-0.1	-2.48	-1.30	-0.09	-2.48	-1.30
		100	-0.00	-0.57	0.58	-0.05	-1.85	-0.73	-0.05	-1.85	-0.72
	100	25	0.45	-0.69	0.70	-0.19	-3.06	-1.82	-0.18	-3.06	-1.82
		50	-0.22	-0.63	0.62	-0.10	-2.47	-1.29	-0.10	-2.47	-1.28
		100	-0.12	-0.58	0.57	-0.05	-1.85	-0.73	-0.05	-1.85	-0.73
	200	25	-0.16	-0.69	0.71	-0.19	-3.06	-1.82	-0.19	-3.06	-1.82
		50	-0.11	-0.63	0.62	-0.10	-2.47	-1.29	-0.10	-2.47	-1.29
		100	-0.06	-0.57	0.58	-0.05	-1.85	-0.73	-0.05	-1.85	-0.73
Percent Bias (%)	50	25	0.03	-10.7	12.7	-3.0	-47.1	-33.1	-2.9	-47.0	-33.0
		50	-0.02	-9.7	11.3	-1.6	-38.2	-23.6	-1.6	-38.1	-23.6
		100	0.00	-8.8	10.5	-0.9	-28.5	-13.2	-0.8	-28.4	-13.1
	100	25	0.07	-10.6	12.7	-3.1	-47.1	-33.1	-3.0	-47.1	-33.1
		50	-0.04	-9.7	11.4	-1.7	-38.0	-23.4	-1.6	-38.0	-23.3
		100	-0.02	-8.9	10.4	-0.9	-28.5	-13.3	-0.9	-28.5	-13.3
	200	25	-0.03	-10.6	12.8	-3.2	-47.1	-33.1	-3.1	-47.1	-33.1
		50	-0.02	-9.7	11.3	-1.7	-38.0	-23.4	-1.7	-38.0	-23.4
		100	-0.01	-8.8	10.5	-0.9	-28.5	-13.3	-0.9	-28.5	-13.3
Standard Error	50	25	0.46	1.14	1.14	0.30	0.60	0.63	0.29	0.61	0.63
		50	0.37	0.92	0.92	0.26	0.56	0.58	0.25	0.56	0.58
		100	0.29	0.71	0.71	0.22	0.49	0.50	0.21	0.50	0.51
	100	25	0.32	0.80	0.80	0.21	0.42	0.44	0.20	0.42	0.44
		50	0.26	0.64	0.64	0.18	0.39	0.40	0.18	0.39	0.40
		100	0.21	0.49	0.49	0.15	0.34	0.35	0.15	0.34	0.35
	200	25	0.23	0.56	0.56	0.15	0.29	0.31	0.14	0.29	0.31
		50	0.18	0.45	0.45	0.13	0.27	0.28	0.12	0.27	0.28
		100	0.14	0.34	0.34	0.11	0.24	0.24	0.11	0.24	0.24

Table 4.3 (cont'd)

Root Mean Squared Error	50	25	0.42	1.32	1.32	0.35	3.12	1.93	0.34	3.11	1.93
		50	0.34	1.12	1.11	0.27	2.54	1.42	0.27	2.54	1.42
		100	0.28	0.92	0.92	0.22	1.92	0.89	0.22	1.91	0.88
	100	25	0.29	1.03	1.04	0.28	3.09	1.87	0.27	3.08	1.87
		50	0.24	0.89	0.89	0.21	2.50	1.35	0.20	2.50	1.35
		100	0.19	0.76	0.75	0.16	1.89	0.81	0.16	1.88	0.81
	200	25	0.21	0.87	0.89	0.24	3.07	1.85	0.24	3.07	1.85
		50	0.17	0.77	0.76	0.16	2.48	1.32	0.16	2.48	1.32
		100	0.14	0.67	0.67	0.12	1.87	0.77	0.12	1.87	0.77

Table 4.4 Sensitivity of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites for the continuous outcome when site size is varied if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$.

Measure	Type of Weight	# of sites	Oracle			Naïve			Bootstrap		
			C1	C2	C3	C1	C2	C3	C1	C2	C3
Bias (× 100)	UN	50	0.50	-0.63	0.64	-0.10	-2.55	-1.36	-0.10	-2.54	-1.35
		100	0.08	-0.62	0.65	-0.11	-2.56	-1.37	-0.11	-2.55	-1.36
		200	-0.07	-0.64	0.63	-0.11	-2.57	-1.38	-0.11	-2.56	-1.37
	SS	50	0.25	-0.62	0.63	-0.09	-2.46	-1.28	-0.09	-2.46	-1.28
		100	0.08	-0.62	0.63	-0.10	-2.46	-1.28	-0.10	-2.46	-1.28
		200	-0.03	-0.62	0.63	-0.10	-2.46	-1.28	-0.10	-2.46	-1.28
	WV	50	0.75	-0.67	0.58	-0.15	-2.57	-1.40	-0.14	-2.56	-1.38
		100	0.04	-0.67	0.59	-0.17	-2.58	-1.40	-0.16	-2.57	-1.40
		200	-0.14	-0.69	0.57	-0.17	-2.59	-1.42	-0.17	-2.59	-1.4
	COMP	50	0.19	-0.62	0.63	-0.07	-2.48	-1.30	-0.06	-2.48	-1.30
		100	0.11	-0.61	0.63	-0.07	-2.47	-1.30	-0.07	-2.47	-1.30
		200	-0.00	-0.62	0.62	-0.07	-2.48	-1.30	-0.07	-2.48	-1.30
Percent Bias (%)	UN	50	0.08	-9.7	11.7	-1.7	-39.3	-24.8	-1.6	-39.1	-24.6
		100	0.01	-9.6	11.8	-1.8	-39.4	-24.9	-1.8	-39.3	-24.8
		200	-0.01	-9.8	11.5	-1.8	-39.5	-25.0	-1.8	-39.5	-25.0
	SS	50	0.04	-9.6	11.4	-1.6	-37.9	-23.3	-1.5	-37.9	-23.3
		100	0.01	-9.5	11.5	-1.6	-37.8	-23.3	-1.6	-37.8	-23.2
		200	0.00	-9.6	11.4	-1.6	-37.9	-23.3	-1.6	-37.9	-23.3
	WV	50	0.12	-10.4	10.6	-2.5	-39.6	-25.4	-2.3	-39.4	-25.2
		100	0.01	-10.3	10.8	-2.8	-39.7	-25.5	-2.6	-39.6	-25.4
		200	-0.02	-10.6	10.4	-2.8	-39.9	-25.7	-2.8	-39.8	-25.6
	COMP	50	0.03	-9.5	11.4	-1.1	-38.1	-23.7	-1.0	-38.2	-23.7
		100	0.02	-9.4	11.5	-1.1	-38.1	-23.6	-1.1	-38.1	-23.6
		200	0.00	-9.7	11.3	-1.1	-38.1	-23.7	-1.1	-38.1	-23.7

Table 4.4 (cont'd)

Standard Error	UN	50	0.38	0.95	0.95	0.27	0.57	0.59	0.27	0.59	0.61
		100	0.27	0.66	0.66	0.19	0.39	0.41	0.19	0.41	0.42
		200	0.19	0.47	0.47	0.13	0.28	0.29	0.13	0.29	0.30
	SS	50	0.37	0.92	0.92	0.26	0.56	0.58	0.25	0.57	0.58
		100	0.26	0.64	0.64	0.18	0.39	0.40	0.18	0.39	0.40
		200	0.18	0.45	0.45	0.13	0.27	0.28	0.12	0.27	0.28
	WV	50	0.35	0.88	0.87	0.26	0.56	0.58	0.29	0.62	0.66
		100	0.24	0.61	0.60	0.18	0.39	0.41	0.21	0.44	0.47
		200	0.17	0.43	0.42	0.13	0.27	0.28	0.15	0.32	0.34
	COMP	50	0.35	0.94	0.94	0.25	0.57	0.59	0.25	0.57	0.59
		100	0.25	0.65	0.65	0.17	0.39	0.41	0.17	0.40	0.41
		200	0.17	0.46	0.46	0.12	0.28	0.28	0.12	0.28	0.29
Root Mean Squared Error	UN	50	0.37	1.16	1.16	0.29	2.62	1.50	0.29	2.61	1.49
		100	0.26	0.93	0.94	0.22	2.59	1.44	0.22	2.59	1.43
		200	0.18	0.80	0.79	0.17	2.58	1.41	0.17	2.58	1.41
	SS	50	0.35	1.11	1.11	0.27	2.53	1.41	0.27	2.53	1.41
		100	0.24	0.89	0.90	0.20	2.49	1.34	0.20	2.49	1.34
		200	0.17	0.77	0.77	0.16	2.48	1.31	0.16	2.48	1.31
	WV	50	0.34	1.15	1.09	0.35	2.66	1.57	0.34	2.64	1.54
		100	0.24	0.93	0.88	0.28	2.62	1.49	0.27	2.62	1.48
		200	0.17	0.82	0.73	0.23	2.61	1.46	0.23	2.61	1.46
	COMP	50	0.34	1.13	1.13	0.26	2.54	1.43	0.26	2.55	1.43
		100	0.24	0.90	0.92	0.19	2.51	1.36	0.19	2.51	1.36
		200	0.17	0.77	0.78	0.14	2.49	1.33	0.14	2.49	1.33

In Table 4.3 and Table 4.4, each row within a performance measure represents one scenario that is described in the previous section. The bias of the oracle estimator is all negative in C2 and all positive in C3. This is consistent with equation (4.7) which becomes $E((\widehat{\boldsymbol{\pi}}'\widehat{\boldsymbol{\pi}})^{-1}\boldsymbol{\pi}'\mathbf{U}\boldsymbol{\gamma})$ because there are no measurement errors for the oracle estimator. As the covariance is all negative in our simulations, the sign of the coefficient of the omitted variable dominates the direction of the bias. On the other hand, there are almost no differences on the mean standard errors if we change the sign of the coefficient of the omitted variable for the oracle estimator within each scenario for the fixed site size. This is also the case for the varied site size except those weighted by the within-site variation which is reasonable as the variability has been taken into account.

When Assumption V is violated, for the naïve estimator and the bootstrap estimator, the bias formula in equation (4.7) show that the bias is related to the difference of the scaled covariance between the omitted variable and the principal strata proportions and the scaled variance of the measurement errors. The scalars are the coefficient of the omitted variable and the population principal strata ITT effects, respectively. Because of this difference, in some circumstances the magnitude of the bias of the naïve estimator or the bootstrap estimator is not significantly larger than that of the oracle estimator within the same scenario for the same case. For example, in Table 4.3, when the number of sites is 50 and the site size is 100, the bias of the oracle estimator for C3 is 0.58 whereas the bias of the native estimator is -0.73 whose absolute value is not significantly bigger than that of the oracle estimator.

The mean standard errors of the naïve estimator and the bootstrap estimator are found to be smaller than those of the oracle estimator when there are omitted variables for both the fixed and varied site size, an identical finding as in the performance analysis. This is true because there is more variability in the estimated principal strata proportions than in the true principal strata

proportions for the naïve estimator and the bootstrap estimator, resulting in less variation in the distribution of the estimators. For the root mean squared errors, we find a consistent pattern for the three estimators across the three cases, that is, the larger the number of sites, the larger the number of individuals in each site, the smaller the root mean squared error, the better the estimators.

CHAPTER 5

APPLICATION

Major depressive disorder (MDD) is a common medical disorder and a serious public health concern. More than 300 million people worldwide suffer from depression, corresponding to 4.4% of the global population (World Health Organization, 2017). Women are more likely to live with depression than men with a prevalence of 5.1% and 3.6%, respectively. The economic burden of MDD has increased by 21.5% to $210.5 billion in 2010 compared to that in 2005 (Greenberg et al., 2015). As one of the leading causes of non-fatal health loss, depression contributes to 7.5% of years lived with disability (YLD) (James et al., 2018). Depression is also a leading cause of suicide deaths, costing approximately 800,000 lives every year.

To treat major depressive disorder, a variety of antidepressants are available (Frank et al., 1996). Among these medications, the selective serotonin reuptake inhibitors (SSRIs), serotonin norepinephrine reuptake inhibitors (SNRIs), mirtazapine, and bupropion are often taken as the first-line treatments because they are safe, more tolerable and acceptable, less costly and have less toxicity (Gelenberg et al., 2010). The effectiveness of these antidepressants, between or within medication classes, is similar. However, approximately 50% of the depressed patients fail to respond to the initial treatments and only about 30% achieve full remission (Fava et al., 2003). Lack of remission to the first-line treatments will result in persistent disabilities and impaired quality of life. Thus, next-step treatments after initial failure to boost remission rates are often in demand.

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial is a large-scale, multicenter, multistep, real-world clinical trial of patients with nonpsychotic major depressive disorder. It aims to provide evidence for the optimal choice of the next-step treatments

for those who have an initial failure to the first-line antidepressant. Details of the study design is described elsewhere (Rush et al., 2004). In this chapter, we apply our proposed approaches to the STAR*D data at level 2 to identify the complier average causal effect (CACE) of treatment options when treatment nonadherence exists. Specifically, we focus on the antidepressant medications at level 2 for those who did not obtain a satisfactory response from citalopram (CIT), an SSRI antidepressant at level 1. Level 2 treatments include augmentation to the previous citalopram or switching to a different medication based on their acceptability. For those who chose the medication augmentation strategy, they were randomly assigned to receive one of the two treatment options: sustained-release bupropion (BUP) or buspirone (BUS). For those who chose the medication switching strategy, they were randomly assigned to receive one of the three treatment options: sustained-release BUP, sertraline (SER), or extended-release venlafaxine (VEN). Among these three switching options, we only consider sustained-release BUP and extended-release VEN for comparison in this dissertation because they belong to different medication classes from citalopram which may provide new indications. With the study setting, we define our estimand of interest, CACE, in terms of its following attributes (ICH E9 working group, 2020).

Treatment. We compare treatment options within the medication augmentation and switching strategies separately. For the medication augmentation strategy, the treatments of interest are augmenting CIT with BUP (CIT+BUP) and augmenting CIT with BUS (CIT+BUS). For the medication switching strategy, the treatments of interest are BUP and VEN.

Population. Our population of interest consists of those who always adhere to whichever treatment they would be assigned regardless of the assignment status, that is, the compliers defined by the potential treatment received under the principal stratification framework in Chapter 3.

Endpoint. The primary endpoint of interest is the remission of symptoms which is defined as a score of seven or less on the 17-item Hamilton Rating Scale for Depression (HAM-D_{17}) at the end (12 weeks or longer) of level 2. We also consider secondary endpoints that are based on the HAM-D_{17} scores. They are 1) total HAM-D_{17} scores at the end of level 2 which range from 0 to 52 with higher values indicating more severe depression; 2) reduction of HAM-D_{17} scores, i.e., (HAM-D_{17} scores at the beginning of level 2 − HAM-D_{17} scores at the end of level 2); and 3) response which is defined as a 50% or more reduction of HAM-D_{17} scores at the end of level 2.

Population-level summary. We use the potential outcome means under each treatment in our population of interest as the population-level summaries.

Intercurrent Event. We are interested in the intercurrent event of treatment nonadherence in this study. Specifically, we consider a wide range of reasons for treatment nonadherence which are collected by the protocol deviation questionnaire and the study termination questionnaire. The reasons include 1) lack of efficacy; 2) unacceptable side effects; 3) committed suicide/suicide attempt; 4) developed general medical or surgical condition that required protocol to be stopped; 5) developed symptoms requiring non-protocol treatment; 6) moved from the area; 7) found research too burdensome; 8) patient nonadherence to study procedures; and 9) patient nonadherence to study medications.

With the above attributes, we can clearly define our causal estimand of interest as the mean difference at week 12 or longer between the potential outcomes under each of the two active treatments for the principal stratum of compliers when treatment nonadherence exists. The corresponding treatment effect is therefore the CACE and we use the corresponding principal stratum strategy in the ICH E9 (R1) addendum to address the intercurrent event of treatment nonadherence.

5.1 Medication Augmentation: CIT+BUP vs. CIT+BUS

In STAR*D, randomization was stratified by clinical sites and preference stratum. After the initial failure of citalopram at level 1, 565 patients in 38 clinical sites accepted the medication augmentation strategy at level 2. Among these patients, 279 were randomly assigned to CIT+BUP and 286 were randomly assigned to CIT+BUS. Table 5.1 shows the characteristics of the patients at baseline or level 2 entry, all of which are balanced between the two treatment assignment groups except the length of illness which is defined as years from first episode to baseline of the study. Those assigned to CIT+BUP had a significantly shorter length of illness (15.2 years) than those assigned to CIT+BUS (17.7 years, $P = 0.02$). Table 5.2 shows the observed treatment adherence by treatment assignment groups. Note that the treatment assignment Z takes two values with 1 representing those assigned to CIT+BUP ($Z = 1$) and 2 representing those assigned to CIT+BUS ($Z = 2$). The treatment receipt A takes three values with 1 representing receiving CIT+BUP ($A = 1$), 2 representing receiving CIT+BUS ($A = 2$) and 0 representing not taking any treatment as required. We observe that 80% of the patients in the CIT+BUP assignment group adhered to what was assigned whereas in the CIT+BUS group 73% of the patients adhered to what was assigned. None of the patients in either group crossed over to receive treatment from the other group.

Table 5.1 Baseline and level 2 entry characteristics of the patients in the medication augmentation strategy.

Characteristic	CIT+BUP (N=279)	CIT+BUS (N=286)	Total (N=565)
Setting			
Primary care	94 (33.7)	95 (33.2)	189 (33.5)
Specialty care	185 (66.3)	191 (66.8)	376 (66.5)
QIDS-C16 at level 2 entry	11.7±4.1	11.7±4.2	11.7±4.2
Monthly household income	2287±2600	2366±3320	2327±2983
Marital status			
Never married	84 (30.1)	79 (27.7)	163 (28.9)
Married or cohabiting	112 (40.1)	117 (41.1)	229 (40.6)
Separated or divorced	73 (26.2)	85 (29.8)	158 (28.0)
Widowed	10 (3.6)	4 (1.4)	14 (2.5)
Education			
Less than high school	33 (11.8)	44 (15.4)	77 (13.7)
Less than college	178 (63.8)	183 (64.2)	361 (64.0)
College or above	68 (24.4)	58 (20.4)	126 (22.3)
Employment status			
Unemployed	111 (39.8)	125 (43.9)	236 (41.9)
Employed	160 (57.3)	147 (51.6)	307 (54.4)
Retired	8 (2.9)	13 (4.5)	21 (3.7)
Insurance			
Private insurance	130 (48.1)	142 (51.6)	272 (49.9)
Public insurance	35 (13.0)	36 (13.1)	71 (13.0)
No insurance	105 (38.9)	97 (35.3)	202 (37.1)
Sex			
Male	107 (38.4)	126 (44.1)	233 (41.2)
Female	172 (61.6)	160 (55.9)	332 (58.8)
Age	41.3±12.8	42.0±12.6	41.6±12.7
Race			
White	221 (79.2)	220 (76.9)	441 (78.1)
Black/African American	48 (17.2)	47 (16.4)	95 (16.8)
Other	10 (3.6)	19 (6.7)	29 (5.1)
Hispanic			
No	243 (87.1)	245 (85.7)	488 (86.4)
Yes	36 (12.9)	41 (14.3)	77 (13.6)
CIRS score	4.5±4.1	4.9±4.0	4.7±4.0
Age at first MDE	26.1±14.4	24.3±13.6	25.2±14.0
Length of illness	15.2±12.6	17.7±13.6	16.5±13.2
Family history of MDD			
No	131 (48.0)	138 (48.6)	269 (48.3)
Yes	142 (52.0)	146 (51.4)	288 (51.7)

Table 5.1 (cont'd)

Suicide history			
No	266 (97.4)	274 (96.8)	540 (97.1)
Yes	7 (2.6)	9 (3.2)	16 (2.9)
SF-12 score			
Physical	48.3±12.5	48.1±12.0	48.2±12.2
Mental	25.0±8.2	26.2±7.7	25.6±8.0
WSAS score	25.8±8.1	25.6±9.1	25.7±8.6
QLESQ score	37.1±12.9	39.3±14.2	38.2±13.6

5.1.1 Assessment of Assumptions

In this section, we discuss the validity of the key assumptions used to identify CACE, that is, Assumptions I, II, and III for the nonparametric approach and Assumption V for the multisite design approach. Specifically, we evaluate whether these assumptions can be falsified for the augmentation strategy. For the nonparametric approach, it is straightforward that Assumption II, the monotonicity assumption, is satisfied because no crossover occurs in our data as shown in Table 5.2.

Table 5.2 Observed treatment adherence in the medication augmentation strategy.

		Treatment Received (A)		
		None ($A = 0$)	CIT+BUP ($A = 1$)	CIT+BUS ($A = 2$)
Treatment Assigned (Z)	CIT+BUP ($Z = 1$)	56 (20)	223 (80)	0 (0)
	CIT+BUS ($Z = 2$)	78 (27)	0 (0)	208 (73)

If the treatment assignment has an effect on the outcomes of interest that is not through the treatment received, Assumption I, the exclusion restriction assumption will be violated. However, the exclusion restriction assumption can never be verified from the data because the potential outcomes cannot be observed at the same time. It can, on the other hand, be testable using available data (Angrist et al., 1996). To better represent the real world, STAR*D was designed not to blind the patients and the physicians who administered the drugs, but only blind the assessors (Sinyor et al., 2010). It is likely that knowing the assigned treatment will have an effect on "experiencing" side effects and subsequently violate the exclusion restriction assumption. Table 5.3 shows the side effects in level 2 in the medication augmentation strategy. No significant difference was found

between those assigned to CIT+BUP and those assigned to CIT+BUS in terms of maximal frequency ($\chi^2 = 0.87, P = 0.99$), maximal intensity ($\chi^2 = 2.04, P = 0.92$), or maximal burden of side effects ($\chi^2 = 12.0, P = 0.06$). Therefore, our data do not support the violation of Assumption I.

Table 5.3 Side effects in level 2 in the medication augmentation strategy.

Characteristic	CIT+BUP (N=279)	CIT+BUS (N=286)	Total (N=565)
Maximal frequency of side effects in level 2			
No side effects	55 (19.8)	59 (20.7)	114 (20.2)
10% of the time	44 (15.8)	47 (16.5)	91 (16.2)
25% of the time	47 (16.9)	49 (17.2)	96 (17.0)
50% of the time	47 (16.9)	52 (18.3)	99 (17.6)
75% of the time	31 (11.2)	28 (9.8)	59 (10.5)
90% of the time	19 (6.8)	16 (5.6)	35 (6.2)
All the time	35 (12.6)	35 (11.9)	69 (12.3)
Maximal intensity of side effects in level 2			
No side effects	55 (19.8)	58 (20.4)	113 (20.0)
Trivial	27 (9.7)	29 (10.2)	56 (10.0)
Mild	55 (19.8)	52 (18.3)	107 (19.0)
Moderate	61 (22.0)	60 (21.0)	121 (21.5)
Marked	44 (15.8)	52 (18.3)	96 (17.0)
Severe	27 (9.7)	29 (10.2)	56 (10.0)
Intolerable	9 (3.2)	5 (1.7)	14 (2.5)
Maximal burden of side effects in level 2			
No side effects	63 (22.7)	75 (26.3)	138 (24.5)
Minimal impairment	62 (22.3)	56 (19.7)	118 (21.0)
Mild impairment	65 (23.4)	52 (18.2)	117 (20.8)
Moderate impairment	50 (18.0)	55 (19.3)	105 (18.6)
Marked impairment	24 (8.6)	33 (11.6)	57 (10.1)
Severe impairment	8 (2.9)	14 (4.9)	22 (3.9)
Unable to function	6 (2.1)	0 (0.0)	6 (1.1)

In addition to satisfying Assumption II, the fact that the treatments are not crossed over in the medication augmentation strategy also provides evidence of the nonexistence of always-1-takers (strata 5) and always-2-takers (strata 9). Given this fact and that Assumption I is likely to hold, the population structure would become Table 5.4 if Assumption III, the no partial-compliers assumption is also satisfied for the nonparametric approach. We can therefore falsify Assumption III using the observed data.

Table 5.4 Population structure when Assumptions I (exclusion restriction), Assumption II (monotonicity), and Assumption III (no partial-compliers) are satisfied for the nonparametric approach.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	1. Never-taker (NT) π_1		
	A(1)=1			6. Complier (COMP) π_6
	A(1)=2			

Specifically, following the discussion in section 4.2, we can estimate the proportion of never-takers π_1 in two ways. One is using the probability of those who were assigned to CIT+BUP but did not take it, i.e., $\Pr(A = 0|Z = 1)$ for π_1 and the other is using the probability of those who were assigned to CIT+BUS but did not take it, i.e., $\Pr(A = 0|Z = 2)$. Similarly, we can use the probability of those who were assigned to CIT+BUP and took it, i.e., $\Pr(A = 1|Z = 1)$ or the probability of those who were assigned to CIT+BUS and took it, i.e., $\Pr(A = 2|Z = 2)$ to estimate

the proportion of compliers π_6. If Assumption III holds, we expect that the two estimated probabilities for π_1 to be equal and that the two estimated probabilities for π_6 to be equal. That is,

$$\hat{\pi}_1 = \widehat{\Pr}(A = 0|Z = 1) = \widehat{\Pr}(A = 0|Z = 2)$$

$$\hat{\pi}_6 = \widehat{\Pr}(A = 1|Z = 1) = \widehat{\Pr}(A = 2|Z = 2) \quad (5.1)$$

However, as shown in Table 5.2, the probability $\widehat{\Pr}(A = 0|Z = 1) = 0.2$ differs significantly from the probability $\widehat{\Pr}(A = 0|Z = 2) = 0.27$, and the probability $\widehat{\Pr}(A = 1|Z = 1) = 0.8$ differs significantly from the probability $\widehat{\Pr}(A = 2|Z = 2) = 0.73$ ($z = -2.01, P = 0.04$). Therefore, there is a possibility that Assumption III is violated. We will give further discussions later in this chapter.

Table 5.5 Population structure when Assumptions I (exclusion restriction), Assumption II (monotonicity), and IV (no never-takers) are satisfied for the multisite design approach.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0			3. Partial-2-complier (P2C) π_3
	A(1)=1	4. Partial-1-complier (P1C) π_4		6. Complier (COMP) π_6
	A(1)=2			

For the multisite design approach, the population structure becomes Table 5.5 under Assumptions I, II, and IV in the medication augmentation strategy. We only evaluate Assumption V, the zero correlation assumption as it is the key to this approach. Specifically, Assumption V can be violated if there are site-level covariates that correlate to both the site-specific principal

strata proportions $\pi_{3|k}$, $\pi_{4|k}$, $\pi_{6|k}$ and the site-level ITT effects $ITT_{y|k}$ for $k = 1, 2, \dots, 38$. As we cannot observe $\pi_{3|k}$, $\pi_{4|k}$, $\pi_{6|k}$ and $ITT_{y|k}$ directly, we will use the estimated quantities $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$, and $\widehat{ITT}_{y|k}$ instead. The site-level variable of clinical setting which indicates whether a site provides primary care or specialty care could be one such site-level covariate. Patients in the specialty care setting might be more compliant to their treatment and therefore have a larger proportion of compliers. However, our data do not provide evidence of association between clinical setting and $\hat{\pi}_{3|k}$ $(t = 0.7, P = 0.49)$, $\hat{\pi}_{4|k}$ $(t = -0.18, P = 0.86)$, $\hat{\pi}_{6|k}$ $(t = -0.29, P = 0.77)$. We also consider aggregated variables from the individual-level data as the potential confounders including site-specific proportion of females, proportion of African Americans, average age, proportion of the unemployed, average household income, average years of schooling, proportion of married individuals, proportion of those with a family history with MDD, average physical health, average mental health, average quality of life, and average baseline medical conditions.

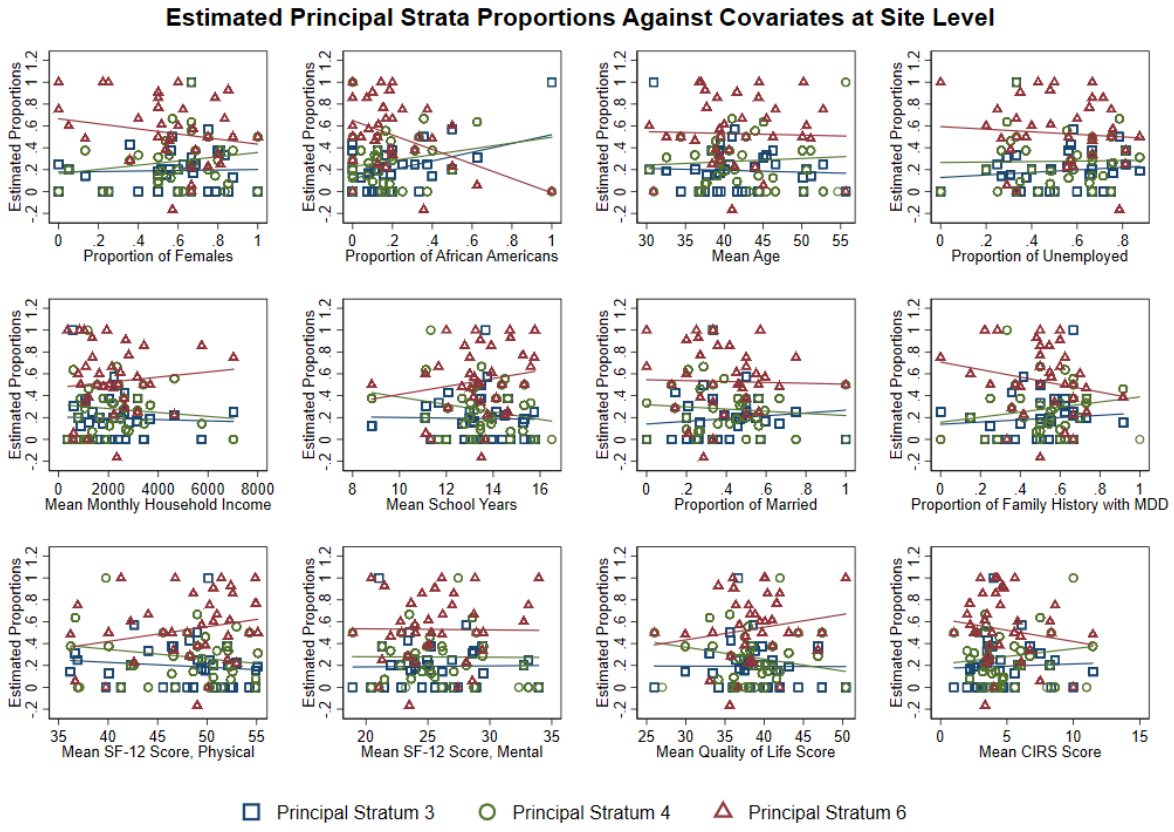

Figure 5.1 Linear regressions of site-specific estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on the site-level aggregated covariates in the medication augmentation strategy.

Figure 5.1 shows the linear regressions of the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on each of the above aggregated variables. No significant association has been found for these covariates except the proportion of African Americans ($\hat{\pi}_{3|k}: t = 4.98, P < 0.001$; $\hat{\pi}_{4|k}: t = 0.2, P = 0.85$; $\hat{\pi}_{6|k}: t = -2.8, P = 0.008$). We further evaluate the association between the site-level $\widehat{ITT}_{y|k}$ of the four endpoints and proportion of African Americans through linear regressions and no significant association has been found for any of the endpoints. Therefore, our data seem to support the plausibility of Assumption V.

5.1.2 Analysis

Our primary and secondary endpoints are all based on the HAM-D_{17} scores at the end of level 2 which have some missing data. Table 5.6 summarizes the reasons for missingness in the medication augmentation strategy. Overall, 61.3% of the patients missed the HAM-D_{17} scores due to loss to follow up or withdrawn consent. 19.1% of the patients are due to treatment nonadherence. 32 (18.5%) patients exited level 2 before week 12. Even though the HAM-D_{17} scores of the last group were collected, we ignore these scores and treat them as missing because otherwise it will misalign with our causal estimand of interest that targets endpoints at week 12 or longer. We name the reason of this type of missingness as "inadequate stay in level 2". There were also 2 (1.1%) patients who missed the HAM-D_{17} scores with no reasons.

Table 5.6 Reasons for missing HAM-D_{17} scores in the medication augmentation strategy.

Reasons for missing HAM-D_{17} scores	CIT+BUP (N=85)	CIT+BUS (N=88)	Total (N=173)
Lost to follow up/Withdrew consent	59 (69.4)	47 (53.4)	106 (61.3)
Inadequate stay in level 2	10 (11.8)	22 (25.0)	32 (18.5)
Treatment nonadherence	15 (17.6)	18 (20.5)	33 (19.1)
Other	1 (1.2)	1 (1.1)	2 (1.1)

To deal with missingness, as suggested in the National Research Council (NRC) report, we chose the most commonly used mixed effects model repeated measures (MMRM) analysis which assumes the missing endpoints are missing at random (MAR) as our main analysis following the framework in Figure 1 of the ICH E9 (R1) addendum. (National Research Council, 2010; ICH E9 working group, 2020) We included age, race, duration of level 2 treatment, and the 16-item Quick Inventory of Depressive Symptomatology-Clinician-Rated (QIDS-C_{16}) scores at the beginning of level 2 in the MMRM analysis because these covariates differed significantly between those with and without missing endpoints. Next, we applied our proposed approaches within the MMRM analysis. Specifically, we estimated the potential outcome $Y_i(1, A(1))$ for each patient in the sample by fitting a random effect model using patients who were randomly assigned to $Z = 1$ only. Then, we obtained $E(Y_i(1, A(1))$ by taking average of the fitted values. Similarly, we estimated $E(Y_i(2, A(2))$ by taking average of the fitted values from a separate random effect model using patients who were randomly assigned to $Z = 2$ only. The marginal ITT effects of Z on Y for the nonparametric approach were calculated by subtracting the estimated $E(Y_i(1, A(1))$ from the estimated $E(Y_i(2, A(2))$. For the multisite design approach, we obtained the site-level marginal ITT effects of Z on Y by repeating the above steps for each site. We estimated the ITT effects of Z on A from the average values A for each value of Z for the nonparametric approach and the site-

specific principal strata proportions from the observed proportions of A given Z for the multisite design approach. We bootstrapped 500 times to obtain the standard errors of our estimates.

The main analysis has the potential to overestimate the missing endpoints because the underlying MAR mechanism assumes that the missing endpoints and non-missing endpoints follow the same distribution. For placebo-controlled trials, overestimation of the missing endpoints will often result in exaggerated treatment effects (Mehrotra, 2019). To address this problem, Mehrotra et al. (2017) proposed an approach that was less in favor of the treatment group for the placebo-controlled trials and therefore provided an estimator that was less likely towards the direction of treatment efficacy. Their basic idea was to use the estimated overall mean of the placebo group from MMRM as the mean of the dropouts in the treatment group. Subsequently, they estimated the overall mean of the treatment group as a weighted sum of the estimated mean of the completers from MMRM and the estimated mean of the dropouts. The weights are the proportions of completers and dropouts in the treatment group respectively.

To evaluate the robustness of our estimators from the main analysis, we extend the approach in Mehrotra et al. (2017) to our setting with two active treatments and conduct three sensitivity analyses. Specifically, let $\hat{\mu}_1$ be the estimated overall mean in the CIT+BUP group obtained from MMRM discussed above, $\hat{\mu}_2$ be the estimated overall mean from MMRM in the CIT+BUS group, $p_{1,miss}$ be the proportion of missing endpoints in the CIT+BUP group, $p_{2,miss}$ be the proportion of missing endpoints in the CIT+BUS group, and $\hat{\tau}_j$ be the estimated ITT effect of Z on Y for sensitivity analysis $j, j = 1, 2, 3$. Sensitivity analysis 1 uses CIT+BUP as the reference group and re-estimates the overall mean of the CIT+BUS group, that is,

$$\hat{\mu}_{2,new} = \left(1 - p_{2,miss}\right)\hat{\mu}_2 + p_{2,miss}\hat{\mu}_1 \tag{5.2}$$

$$\hat{\tau}_1 = \hat{\mu}_{2,new} - \hat{\mu}_1 = \left(1 - p_{2,miss}\right)(\hat{\mu}_2 - \hat{\mu}_1)$$

Similarly, sensitivity analysis 2 uses CIT+BUS as the reference group and re-estimates the overall mean of the CIT+BUP group, that is,

$$\hat{\mu}_{1,new} = \left(1 - p_{1,miss}\right)\hat{\mu}_1 + p_{1,miss}\hat{\mu}_2 \tag{5.3}$$

$$\hat{\tau}_2 = \hat{\mu}_2 - \hat{\mu}_{1,new} = \left(1 - p_{1,miss}\right)(\hat{\mu}_2 - \hat{\mu}_1)$$

Sensitivity analysis 3 cross-references one group for the other and re-estimates the overall means for both groups, that is,

$$\hat{\mu}_{1,new} = \left(1 - p_{1,miss}\right)\hat{\mu}_1 + p_{1,miss}\hat{\mu}_2 \tag{5.4}$$

$$\hat{\mu}_{2,new} = \left(1 - p_{2,miss}\right)\hat{\mu}_2 + p_{2,miss}\hat{\mu}_1$$

$$\hat{\tau}_3 = \hat{\mu}_{2,new} - \hat{\mu}_{1,new} = \left(1 - p_{1,miss} - p_{2,miss}\right)(\hat{\mu}_2 - \hat{\mu}_1)$$

Note that $\hat{\mu}_2 - \hat{\mu}_1$ is the estimated ITT effect of Z on Y for the main analysis.

5.1.3 Results

Figure 5.2 shows the estimated CACEs and their 95% confidence intervals for the primary endpoint of remission and the secondary endpoints of total HAM-D_{17} scores, reduction of HAM-D_{17} scores, and response. In each panel, the solid horizontal line represents the null CACE. The dashed vertical line separates the nonparametric estimates on the left-hand side from the multisite design estimates on the right-hand side. For the multisite design approach, we list the results with different types of weights. For remission, all of the estimated CACEs are around zero and the 95% confidence intervals include zero; thus no significant differences of the two augmented medications are found in the main analysis and the sensitivity analyses. This aligns with existing findings of similar remission rates for these two groups although the target populations are not the same (Trivedi et al., 2006). Evidently, either augmenting CIT with BUP or augmenting CIT with BUS in the medication augmentation strategy is no better than the other one on remission.

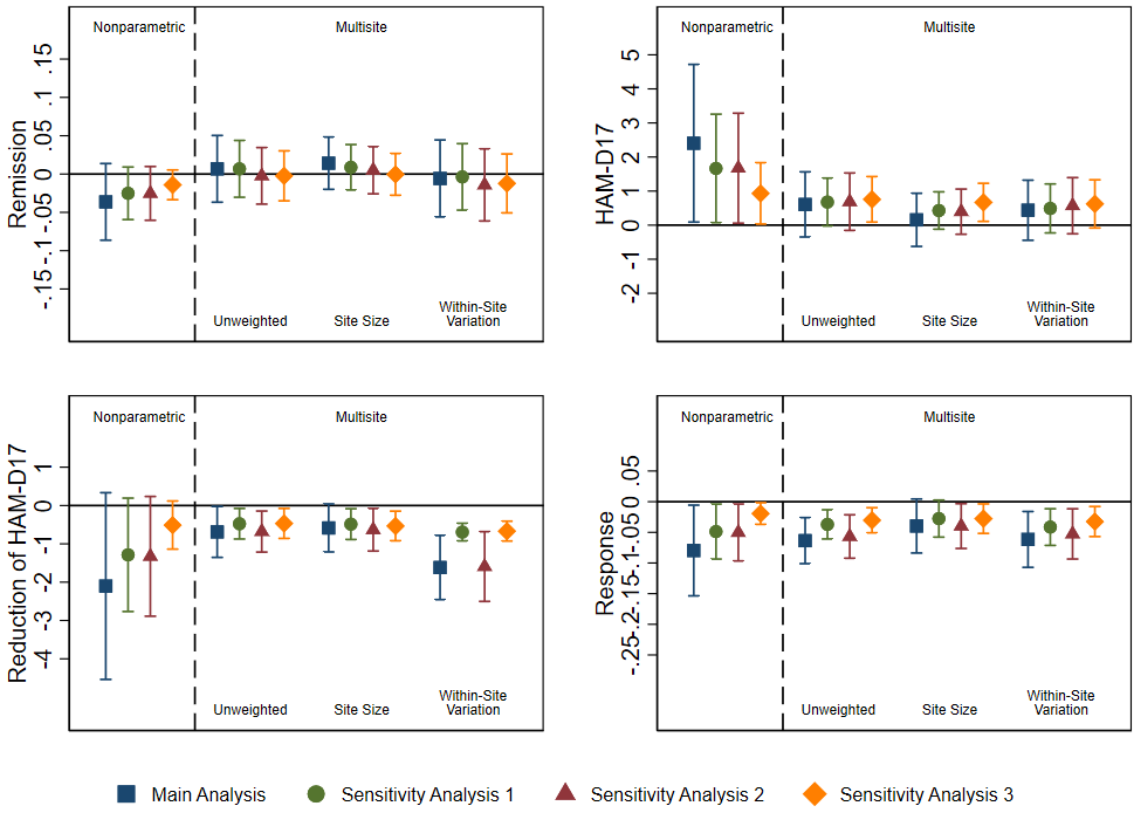

Figure 5.2 Estimated complier average causal effects (CACEs) and their 95% confidence intervals for the primary endpoint and secondary endpoints in the medication augmentation strategy. In each panel, the horizontal solid line represents the null. The dashed vertical line separates the nonparametric estimates on the left-hand side from the multisite design estimates on the right-hand side. The blue square represents main analysis. The green circle represents sensitivity analysis 1. The red triangle represents sensitivity analysis 2. The orange diamond represents sensitivity analysis 3.

The nonparametric approach provides significant estimates of CACE on the total HAM-D_{17} scores at the end of level 2 as shown in the upper right panel of Figure 5.2. As larger HAM-D_{17} scores indicate greater severity of depression, the positive estimated CACEs imply that the symptoms of those augmenting CIT with BUP are less severe at the end of level 2 than those augmenting CIT with BUS. This is also true for the multisite design approach, although no evidence of statistical significance is found except in sensitivity analysis 3 if unweighted or weighted by within-site variation. Given the definition of the reduction of HAM-D_{17} scores in the previous section, we are confident that a greater reduction of symptoms occurs in those augmenting CIT with BUP than those augmenting CIT with BUS because of the negative estimated CACEs across all analyses which is shown in the lower left panel of Figure 5.2. Besides, the multisite design approach also shows that the estimated CACEs are significant except in the main analysis if unweighted or weighted by site size. The estimated CACEs on response are significant in both the nonparametric approach and multisite design approach. In other words, the negative estimated CACEs indicate that those augmenting CIT with BUP have significantly higher response rates than those augmenting CIT with BUS.

Our findings in the main analysis are robust and can be justified from the three sensitivity analyses. In addition, Figure 5.2 also shows that the absolute values of the estimated CACEs in the main analysis are the largest and that those in sensitivity analysis 3 are the smallest on any endpoint in the nonparametric approach. This is as expected because the nonparametric estimates are ratios of the estimated ITT effects of Z on Y and ITT effects of Z on A, and the estimated ITT effects of Z on Y in sensitivity analyses are a portion of the estimated ITT effects of Z on Y in the main analysis as shown in equations (5.2), (5.3), or (5.4). The standard errors in the main analysis are always larger than those in sensitivity analysis 3 in both approaches. This is reasonable because

in sensitivity analysis 3 the estimated mean of the endpoints in either group is a mixture of the non-missing endpoints from both treatment groups, which could reduce the variations. The estimated CACEs of the reduction of HAM-D_{17} scores in the multisite design approach if weighted by within-site variation may look odd. This is because extreme weight is placed on one particular site whose proportion of missingness in treatment assignment group 1 ($Z = 1$) happen to be zero. Therefore, the estimated ITT effects of Z on Y in sensitivity analysis 2 are the same as the main analysis, and the absolute values of the estimated ITT effects of Z on Y in the main analysis and sensitivity analysis 2 are larger than those in sensitivity analysis 1 and sensitivity analysis 3.

5.2 Medication Switching: BUP vs. VEN

For the 727 patients who failed to remit from citalopram and who accepted the medication switching strategy, 239 were randomly assigned to sustained-release bupropion (BUP), 238 were randomly assigned to sertraline (SER), and 250 were randomly assigned to extended-release venlafaxine (VEN). As we aim to find the optimal second-step treatment and sertraline falls into the same medication class SSRI as citalopram, we decided to only compare BUP and VEN for our subsequent analysis. As a result, our final sample contains 489 patients in 38 clinical sites. Table 5.7 shows the characteristics of the patients at baseline or level 2 entry in the medication switching strategy.

Table 5.7 Baseline and level 2 entry characteristics of the patients in the medication switching strategy.

Characteristic	BUP (N=239)	VEN (N=250)	Total (N=489)
Setting			
Primary care	94 (39.3)	96 (38.4)	190 (38.9)
Specialty care	145 (60.7)	154 (61.6)	299 (61.1)
QIDS-C16 at level 2 entry	14.1±4.6	14.1±4.6	14.1±4.6
Monthly household income	2139±2737	2077±2258	2108±2508
Marital status			
Never married	69 (28.9)	63 (25.2)	132 (27.0)
Married or cohabiting	89 (37.2)	104 (41.6)	193 (39.5)
Separated or divorced	69 (28.9)	72 (28.8)	141 (28.8)
Widowed	12 (5.0)	11 (4.4)	23 (4.7)
Education			
Less than high school	25 (10.5)	27 (10.8)	52 (10.6)
Less than college	157 (65.7)	165 (66.0)	322 (65.9)
College or above	57 (23.8)	58 (23.2)	115 (23.5)
Employment status			
Unemployed	90 (37.8)	100 (40.0)	190 (38.9)
Employed	132 (55.5)	138 (55.2)	270 (55.3)
Retired	16 (6.7)	12 (4.8)	28 (5.8)
Insurance			
Private insurance	106 (44.5)	113 (46.9)	219 (45.7)
Public insurance	45 (18.9)	33 (13.7)	78 (16.3)
No insurance	87 (36.6)	95 (39.4)	182 (38.0)
Sex			
Male	103 (43.1)	90 (36.0)	193 (39.5)
Female	136 (56.9)	160 (64.0)	296 (60.5)
Age	42.4±12.9	41.6±12.7	42.0±12.8
Race			
White	179 (74.9)	186 (74.4)	365 (74.6)
Black/African American	47 (19.7)	42 (16.8)	89 (18.2)
Other	13 (5.4)	22 (8.8)	35 (7.2)
Hispanic			
No	216 (90.4)	221 (88.4)	437 (89.4)
Yes	23 (9.6)	29 (11.6)	52 (10.6)
CIRS score	5.1±4.2	5.2±4.1	5.2±4.2
Age at first MDE	25.7±14.6	24.4±13.8	25.0±14.2
Length of illness	16.6±13.6	17.3±13.6	17.0±13.6
Family history of MDD			
No	115 (48.5)	112 (45.7)	227 (47.1)
Yes	122 (51.5)	133 (54.3)	255 (52.9)

Table 5.7 (cont'd)

Suicide history			
No	226 (95.4)	237 (96.7)	463 (96.1)
Yes	11 (4.6)	8 (3.3)	19 (3.9)
SF-12 score			
Physical	46.7±12.4	47.1±12.3	46.9±12.3
Mental	25.8±8.5	26.0±7.7	25.9±8.1
WSAS score	25.2±8.9	25.3±8.6	25.2±8.7
QLESQ score	39.2±15.7	38.4±15.4	38.8±15.5

None of these characteristics are significantly different between the two treatment assignment groups. Table 5.8 shows the observed treatment adherence by treatment assignment groups for patients who accepted the medication switching strategy. For those assigned to BUP ($Z = 1$), 33% did not take the treatment as required ($A = 0$) and 67% adhered to the treatment ($A = 1$). For those assigned to VEN ($Z = 2$), 27% did not take the treatment as required ($A = 0$) and 73% adhered to the treatment ($A = 2$). Again, given the STAR*D design, none of the patients in each treatment assignment group crossed over to obtain the treatment from the other group.

Table 5.8 Observed treatment adherence in the medication switching strategy.

		Treatment Received (A)		
		None ($A = 0$)	BUP ($A = 1$)	VEN ($A = 2$)
Treatment Assigned (Z)	BUP ($Z = 1$)	79 (33)	160 (67)	0 (0)
	VEN ($Z = 2$)	67 (27)	0 (0)	183 (73)

5.2.1 Assessment of Assumptions

We examine the validity of key assumptions in identifying CACE for the medication switching strategy in this section. For the nonparametric approach, we test whether treatment assignment can affect the outcomes through side effects. Table 5.9 shows the side effects in level 2 in the medication switching strategy. No significant difference was found between those assigned to BUP and those assigned to VEN in terms of maximal frequency ($\chi^2 = 11.2, P = 0.08$), maximal intensity ($\chi^2 = 7.74, P = 0.26$), or maximal burden of side effects ($\chi^2 = 8.5, P = 0.2$). Therefore, Assumption I cannot be falsified. Due to the design and criteria of STAR*D, Assumption II will hold directly as no crossover happens. Given Assumption I and II, if Assumption III holds, two

principal strata, which are the never-takers and the compliers, are left as shown in Table 5.4. We can then test Assumption III, the no partial-compliers assumption as we did in section 5.1.1. Specifically, we expect that the proportion of those assigned to BUP but did not take it did not differ significantly from the proportion of those assigned to VEN but did not take it and that the proportion of those assigned to BUP and took it did not differ significantly from the proportion of those assigned to VEN and took it. Even though $\Pr(A = 0|Z = 1) = 0.33$ is not identical to $\Pr(A = 0|Z = 2) = 0.27$, and $\Pr(A = 1|Z = 1) = 0.67$ is not identical to $\Pr(A = 2|Z = 2) = 0.73$, these differences are not significant ($z = 1.51, P = 0.13$). Therefore, we believe that Assumption III is likely satisfied.

Table 5.9 Side effects in level 2 in the medication switching strategy.

Characteristic	BUP (N=239)	VEN (N=250)	Total (N=489)
Maximal frequency of side effects in level 2			
No side effects	34 (14.2)	17 (6.8)	51 (10.4)
10% of the time	10 (4.2)	15 (6.0)	25 (5.1)
25% of the time	21 (8.8)	35 (14.0)	56 (11.4)
50% of the time	43 (18.0)	38 (15.2)	81 (16.6)
75% of the time	39 (16.3)	40 (16.0)	79 (16.2)
90% of the time	29 (12.1)	31 (12.4)	60 (12.3)
All the time	63 (26.4)	74 (29.6)	137 (28.0)
Maximal intensity of side effects in level 2			
No side effects	31 (13.0)	18 (7.2)	49 (10.0)
Trivial	5 (2.1)	7 (2.8)	12 (2.5)
Mild	21 (8.8)	31 (12.4)	52 (10.6)
Moderate	38 (15.9)	42 (16.8)	80 (16.4)
Marked	61 (25.5)	56 (22.4)	117 (23.9)
Severe	46 (19.2)	60 (24.0)	106 (21.7)
Intolerable	37 (15.5)	36 (14.4)	73 (14.9)
Maximal burden of side effects in level 2			
No side effects	34 (14.2)	25 (10.0)	59 (12.1)
Minimal impairment	27 (11.3)	16 (6.4)	43 (8.8)
Mild impairment	28 (11.7)	39 (15.6)	67 (13.7)
Moderate impairment	47 (19.7)	56 (22.4)	103 (21.0)
Marked impairment	50 (20.9)	64 (25.6)	114 (23.3)
Severe impairment	33 (13.8)	33 (13.2)	66 (13.5)
Unable to function	20 (8.4)	17 (6.8)	37 (7.6)

We test Assumption V for the multisite design approach using the same set of site-level covariates as in section 5.1.1. The site-level covariate of clinical setting is not significantly associated with the site-specific estimated principal strata proportions $\hat{\pi}_{3|k}$ ($t = -1.08, P = 0.29$), $\hat{\pi}_{4|k}$ ($t = -0.87, P = 0.39$), $\hat{\pi}_{6|k}$ ($t = 0.95, P = 0.35$). Figure 5.3 shows the linear regression of the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on each of the aggregated variables. No significant association has been found for these covariates except the proportion of African Americans ($\hat{\pi}_{3|k}: t = 0.69, P = 0.5$; $\hat{\pi}_{4|k}: t = 2.1, P = 0.04$; $\hat{\pi}_{6|k}: t = -2.0, P = 0.05$), the proportion of family history with MDD ($\hat{\pi}_{3|k}: t = 2.92, P = 0.006$; $\hat{\pi}_{4|k}: t = -0.19, P = 0.85$; $\hat{\pi}_{6|k}: t = -1.37, P = 0.18$), and the mean quality of life ($\hat{\pi}_{3|k}: t = -2.62, P = 0.01$; $\hat{\pi}_{4|k}: t = -2.68, P = 0.01$; $\hat{\pi}_{6|k}: t = 2.9, P = 0.006$). We further evaluate the association between the site-level $\widehat{ITT}_{y|k}$ of the four endpoints and these three covariates through linear regressions and no significant association has been found for any of the endpoints. Therefore, neither the site-level variable of clinical setting nor the various aggregated variables confound the associations between the estimated site-specific principal strata proportions and the estimated site-level ITT effects. We cannot falsify Assumption V using our data.

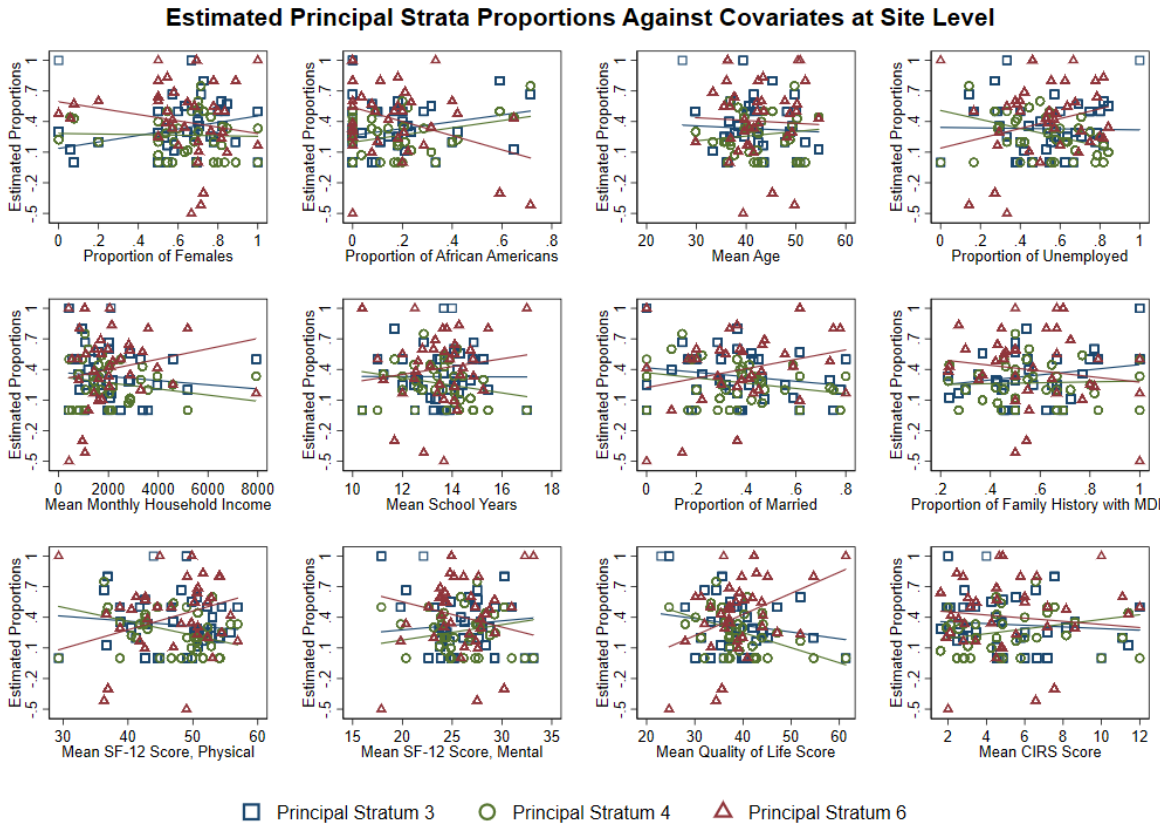

Figure 5.3 Linear regressions of site-specific estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$, $\hat{\pi}_{6|k}$ on the site-level aggregated covariates in the medication switching strategy.

5.2.2 Results

In the medication switching strategy, 186 out of 489 patients who were either randomly assigned to BUP or VEN missed their HAM-D_{17} scores. Out of these patients, 56.5% are due to loss to follow up or withdrawn consent and 20.4% are due to inadequate stay in level 2, as shown in Table 5.10. 39 (21.0%) patients missed their endpoints due to treatment nonadherence. There were 4 (2.1%) patients whose HAM-D_{17} scores were missing with no reasons reported.

Table 5.10 Reasons for missing HAM-D_{17} scores in the medication switching strategy.

Reasons for missing HAM-D_{17} scores	BUP (N=94)	VEN (N=92)	Total (N=186)
Lost to follow up/Withdrew consent	49 (52.1)	56 (60.9)	105 (56.5)
Inadequate stay in level 2	22 (23.4)	16 (17.4)	38 (20.4)
Treatment nonadherence	20 (21.3)	19 (20.6)	39 (21.0)
Other	3 (3.2)	1 (1.1)	4 (2.1)

We also conduct the main analysis and three sensitivity analyses for the medication switching strategy. We use covariates that differ significantly between those with and without missing endpoints in the MMRM analysis. Specifically, we use QIDS-C_{16} scores at the beginning of level 2, education and duration of level 2 treatment for the endpoints of remission and total HAM-D_{17} scores and QIDS-C_{16} scores at the beginning of level 2 and duration of level 2 treatment for the endpoints of change of HAM-D_{17} scores and response. As shown in Figure 5.4, our results provide some clinical implications even though no statistical significance of the estimates of CACE is found on any of the endpoints. The estimated CACEs on remission are all above or almost zero. The total HAM-D_{17} scores at the end of level 2 for those switching to VEN are no bigger than those switching to BUP given the non-positive estimated CACEs. The estimated CACEs on

reduction of HAM-D_{17} scores are non-negative, suggesting that a greater or equal reduction happens after switching to VEN compared to switching to BUP. We do not show the results on response in Figure 5.3 because the random effects models on response do not converge. Our findings are robust in all analyses with main analysis having the widest confidence intervals and sensitivity analysis 3 the narrowest.

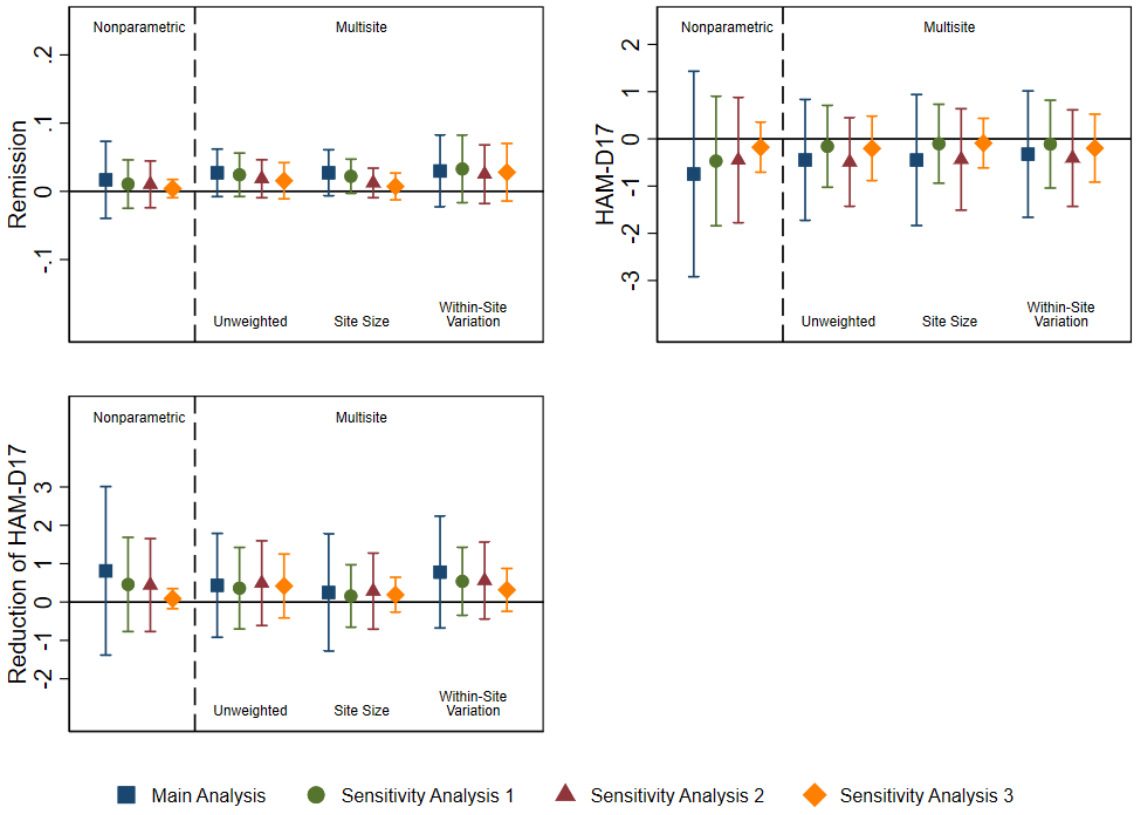

Figure 5.4 Estimated complier average causal effects (CACEs) and their 95% confidence intervals for the primary endpoint and secondary endpoints in the medication switching strategy. In each panel, the horizontal solid line represents the null. The dashed vertical line separates the nonparametric estimates on the left-hand side from the multisite design estimates on the right-hand side. The blue square represents main analysis. The green circle represents sensitivity analysis 1. The red triangle represents sensitivity analysis 2. The orange diamond represents sensitivity analysis 3.

5.3 Discussion

With the observed treatment adherence proportions in each treatment assignment group and the population structure, the population principal strata proportions can be estimated. Specifically, in the medication augmentation strategy, combining Table 5.2 with Table 5.4, 20 to 27 percent of the population are never-takers whereas the rest 73 to 80 percent are compliers for the nonparametric approach. For the multisite design approach in this strategy, with Table 5.2 and Table 5.5, the estimated proportion of partial-1-compliers is 0.27, the estimated proportion of partial-2-compliers is 0.2, and the estimated proportion of compliers is 0.53 ($0.8 - 0.27 = 0.73 - 0.2 = 0.53$). Similarly, in the medication switching strategy, for the nonparametric approach, the estimated proportion of never-takers ranges from 0.27 to 0.33 and the estimated proportion of compliers ranges from 0.67 to 0.73 given Table 5.8 and Table 5.4. For the multisite design in this strategy, the estimated proportions of partial-2-compliers, partial-1-compliers and compliers are 0.33, 0.27 and 0.4 ($0.67 - 0.27 = 0.73 - 0.33 = 0.4$) respectively given Table 5.8 and Table 5.5.

The above proportions are estimated assuming all assumptions required in the nonparametric and multisite design approaches are satisfied. However, as mentioned in section 5.1.1, it is likely that Assumption III, the no partial-compliers assumption, is violated for the nonparametric approach in the medication augmentation strategy. We revisit this problem here. The violation of Assumption III may result from two situations which are shown in Table 5.11 and Table 5.12. In situation 1, the existence of partial-1-compliers violates Assumption III whereas no partial-2-compliers exist. Therefore, given the observed treatment adherence proportions, the estimated principal strata proportions of never-takers, partial-1-compliers and compliers are 0.2, 0.07 and 0.73 respectively.

Table 5.11 Population structure when Assumption III (no partial-compliers) is violated for the nonparametric approach in the medication augmentation strategy: situation 1.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0	$\hat{\pi}_1 = 0.2$		
	A(1)=1	$\hat{\pi}_4 = 0.07$		$\hat{\pi}_6 = 0.73$
	A(1)=2			

Situation 2 has the same population structure as the multisite design approach. Both partial-1-compiers and partial-2-compliers exist in this situation which violates Assumption III. Therefore, the estimated principal strata proportions of partial-2-compliers, partial-1-compliers and compliers are 0.2, 0.27 and 0.53 respectively.

Table 5.12 Population structure when Assumption III (no partial-compliers) is violated for the nonparametric approach in the medication augmentation strategy: situation 2.

		Z=2		
		A(2)=0	A(2)=1	A(2)=2
Z=1	A(1)=0			$\hat{\pi}_3 = 0.2$
	A(1)=1	$\hat{\pi}_4 = 0.27$		$\hat{\pi}_6 = 0.53$
	A(1)=2			

With the estimated proportions of partial-compliers, we can then apply equation (3.9) in chapter 3 to examine the estimators. Specifically, if we assume the ITT effects of Z on Y in partial-compliers are the same as in compliers and plug in the estimated proportions, the nonparametric estimators for situation 1 and situation 2 become,

$$\text{Situation 1: } \frac{E[Y_i(2,A_i(2)) - Y_i(1,A_i(1))]}{E[A_i(2) - A_i(1)]} = CACE + 0.21 * CACE = 1.21 * CACE$$

$$\text{Situation 2: } \frac{E[Y_i(2,A_i(2)) - Y_i(1,A_i(1))]}{E[A_i(2) - A_i(1)]} = CACE + 0.52 * CACE = 1.52 * CACE.$$

It is therefore evident that the nonparametric estimators in both situations overestimate the CACE. Our results also resonate with this evidence. Figure 5.2 shows that our nonparametric estimates for any endpoints are much farther away from the null than the multisite design estimates.

Even though the zero correlation assumptions of the multisite design approach are likely satisfied by our data, the estimators are still biased because of the measurement errors in estimating the principal strata proportions. Specifically, since Assumption V, the zero correlation assumption is likely satisfied by our data, the term $\boldsymbol{\pi}'\mathbf{U}\boldsymbol{\gamma}$ in bias formula (4.7) will not exist. The bias due to measurement errors is then negatively proportional to the true CACE with the proportion less than 1 because it is the ratio of the variance of measurement errors and the variance of the estimated principal strata proportions. With the negative bias, the multisite design approach will therefore underestimate the CACE.

Our analyses provide several meaningful implications. Augmenting citalopram with sustained-release bupropion for remission is no better than augmenting citalopram with buspirone, as indicated by the approximately zero estimated CACE. The estimated CACE on the total HAM-D_{17} score ranges from 0.16 to 2.4 from the main analysis, suggesting that the total HAM-D_{17} score at the end of level 2 if augmenting citalopram with sustained-release bupropion is 0.16 to 2.4 smaller than that if augmenting citalopram with buspirone. Similarly, augmenting citalopram with

sustained-release bupropion reduce the HAM-D_{17} score 0.58 to 2.1 more than augmenting citalopram with buspirone. The response rate if augmenting citalopram with sustained-release bupropion is significantly higher by 4% to 8% than if augmenting citalopram with buspirone. Switching to extended-release venlafaxine and switching to sustained-release bupropion will have similar effects because none of the analyses for any of the four endpoints show significant difference.

Our analyses address the previous criticisms of the STAR*D findings summarized in Pigott (2015). First, corresponding to the first criticism on page 2 of Pigott (2015), instead of using the self-report 16-item Quick Inventory of Depressive Symptomatology (QIDS-SR_{16}) scores that may introduce bias, our primary and secondary endpoints are all based on the HAM-D_{17} scores which are collected by the blinded research outcomes assessors. Second, Pigott (2015) criticized the exclusion of patients with missing exit HAM-D_{17} scores or classification of these patients as nonremitters in previous reports (Rush, Trivedi, Wisniewski, Nierenberg, et al., 2006; Rush, Trivedi, Wisniewski, Stewart, et al., 2006; Trivedi et al., 2006). Our main analysis does not exclude any patients and uses the mixed effects models repeated measures approach for the missing endpoints problem. Our main analysis may be criticized by Pigott as in his third argument because the MMRM method implicitly assumes that the missing endpoints and non-missing endpoints follow the same distribution. We address this criticism by adopting three sensitivity analyses which are described in section 5.1.2. Our results from the sensitivity analyses are consistent with the results in the main analysis.

CHAPTER 6

DISCUSSION AND FUTURE WORK

In this dissertation, we consider treatment nonadherence in randomized controlled trials with two active treatments. We employ the Neyman-Rubin causal framework and the principal stratification framework to explore an alternative causal estimand to the intention to treat (ITT) estimand. With treatment nonadherence, the ITT estimand identifies the causal effect of assignment instead of the causal effect of treatment, which therefore loses clinical relevance. To identify the causal effect of treatment, previous work has largely focused on the placebo-controlled trials. For example, Angrist et al. (1996) identified the complier average causal effects (CACE) under the exclusion restriction assumption and the monotonicity assumption. Yuan et al. (2018) identified the CACE under a multisite design. However, little attention has been placed on randomized controlled trials with two active treatments when treatment nonadherence exists. The identification of the causal estimand becomes complicated in such setting because the population is more diverse than that in the placebo-controlled trial. Therefore, we clarify and define the identification strategies for the CACE.

Our proposed nonparametric approach identifies the complier average causal effect as the ratio of the ITT effect of treatment assignment Z on the outcome Y to the ITT effect of treatment assignment Z on the treatment received A under the exclusion restriction, monotonicity, and no partial-compliers assumptions. We also derive the bias formula and evaluate the performance of the estimator if one of the aforementioned assumptions is violated. Our simulations show that the nonparametric approach can yield a slightly biased estimator for CACE. However, when sample size is 500 or above, the nonparametric estimator becomes unbiased when the percentage of

compliers is above or equal to 70%. In addition, increasing the number of compliers has the potential to reduce the bias to as close as zero when there is deviation from the assumptions.

Our second approach identifies the complier average causal effect via a multisite design under the zero correlation assumption. Under this assumption our simulations show that this approach can yield an unbiased estimator if there are no measurement errors. We derive the bias formula for the situations where measurement errors exist and the zero-correlation assumption is violated. We find that increasing the number of people in each site can reduce the bias because it reduces the variation of the measurement errors. Increasing the number of sites, on the other hand, does not make a significant impact on the bias.

Our study can provide several meaningful insights. First, we identify the causal estimand for the causal effects of treatment, which is clinically relevant and meaningful. Second, we explicitly clarify the underlying assumptions needed to identify our causal estimand of interest. This can help researchers better understand and interpret their results. Third, our derived bias formulas allow researchers to discover potential factors that can result in the bias. Besides, researchers can also improve their study design before conducting any studies based on the bias formulas. Fourth, our results from the nonparametric approach provide evidence for researchers to take measures and increase the adherence rates of participants in a sample. Fifth, our results from the multisite design approach provide evidence for researchers to increase enrollment in each site of the studies. Sixth, our approaches can be extended to any studies that involve a valid instrumental variable and unmeasured confounding and all the above implications can be applied.

Although we make suggestions on increasing the number of compliers based on the nonparametric approach, one limitation of our study is that we cannot identify the complier population in practice. However, we can model the compliance status from observed covariates

and predict an individual's compliance status instead (Roy et al., 2008). Next, one consequence of the non-identifiability of the principal strata is that the measurement errors in estimating the principal strata proportions in our multisite design approach are unavoidable. Therefore, one direction of the future work can be to address the bias issue resulting from the measurement errors. For example, we can substitute a random effects model for the multiple linear regression model used in equation (4.2) to take into account the variations and reduce the bias due to measurement errors (Reardon et al., 2014; Bloom et al., 2017). Another limitation of our study is that the zero correlation assumption used in the multisite design approach is a rather strong assumption and cannot be verified empirically. Future work can be done to incorporate covariates into this assumption and relax it to be a conditional assumption (Yuan et al., 2018). Both approaches in our study make point-identification of the complier average causal effects. To justify the findings, we can also extend our work by introducing bounds for our estimators in the future. Finally, we can extend our study setting to three active treatments, four active treatments, etc. and make pairwise comparisons. For example, in a RCT with three active treatments, the population can be divided into $4 \times 4 = 16$ principal strata based on the potential treatment received. We can then identify the causal estimand from any pair of the three treatments by making additional assumptions.

APPENDICES

APPENDIX A

Results of the Binary Outcomes for the Nonparametric Approach

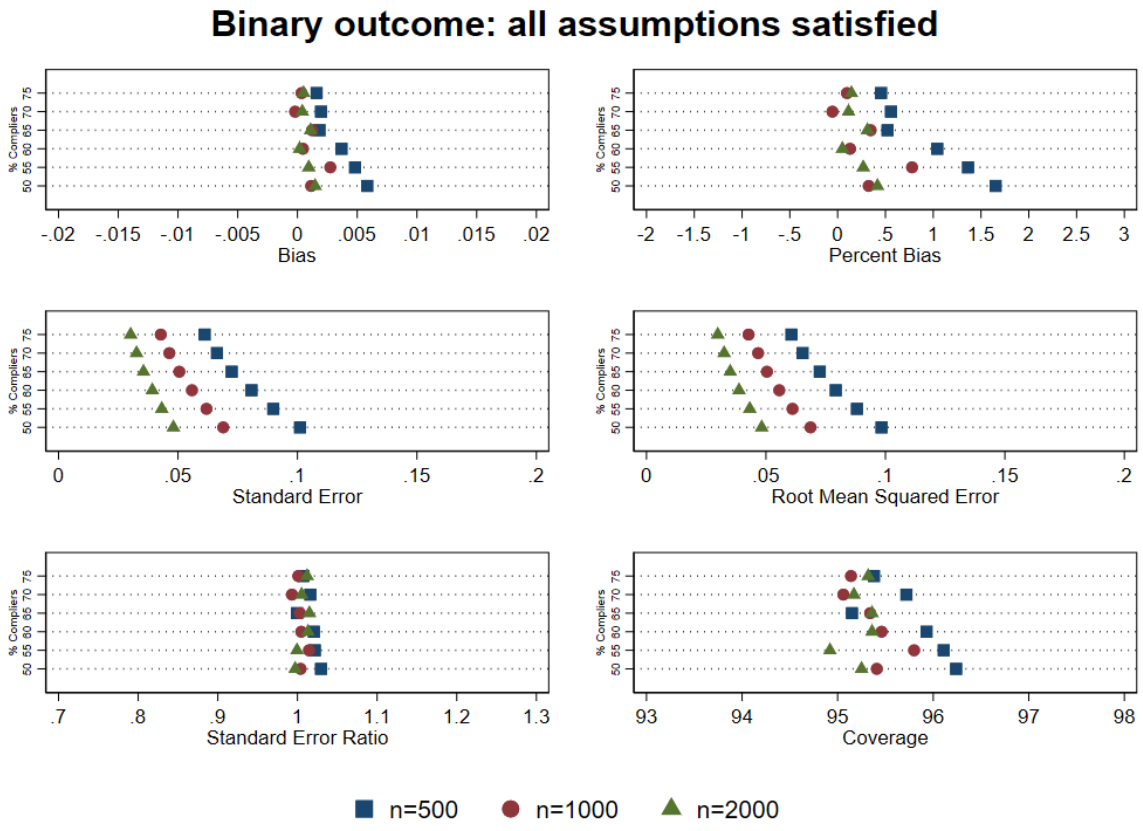

Figure A.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the nonparametric estimator across proportions of compliers for the binary outcome when all assumptions are satisfied (scenario A). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

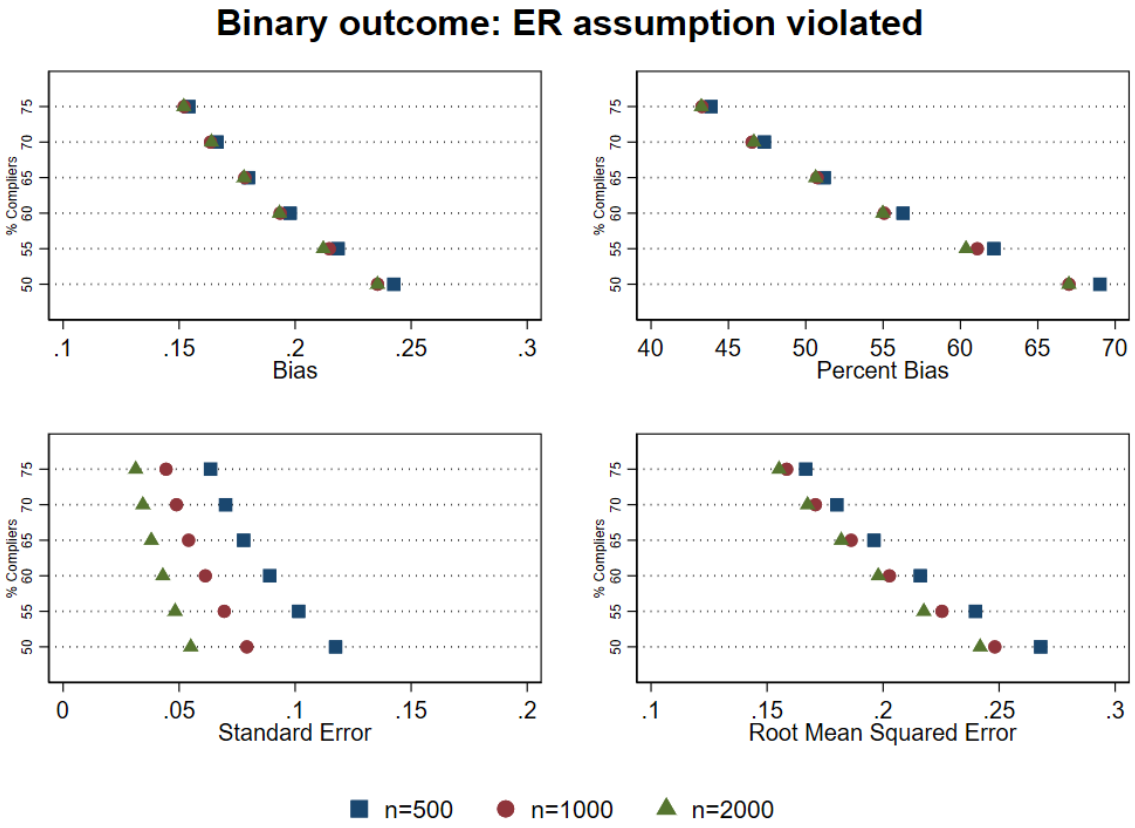

Figure A.2 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of compliers for the binary outcome when Assumption I, the exclusion restriction assumption, is violated (scenario B). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

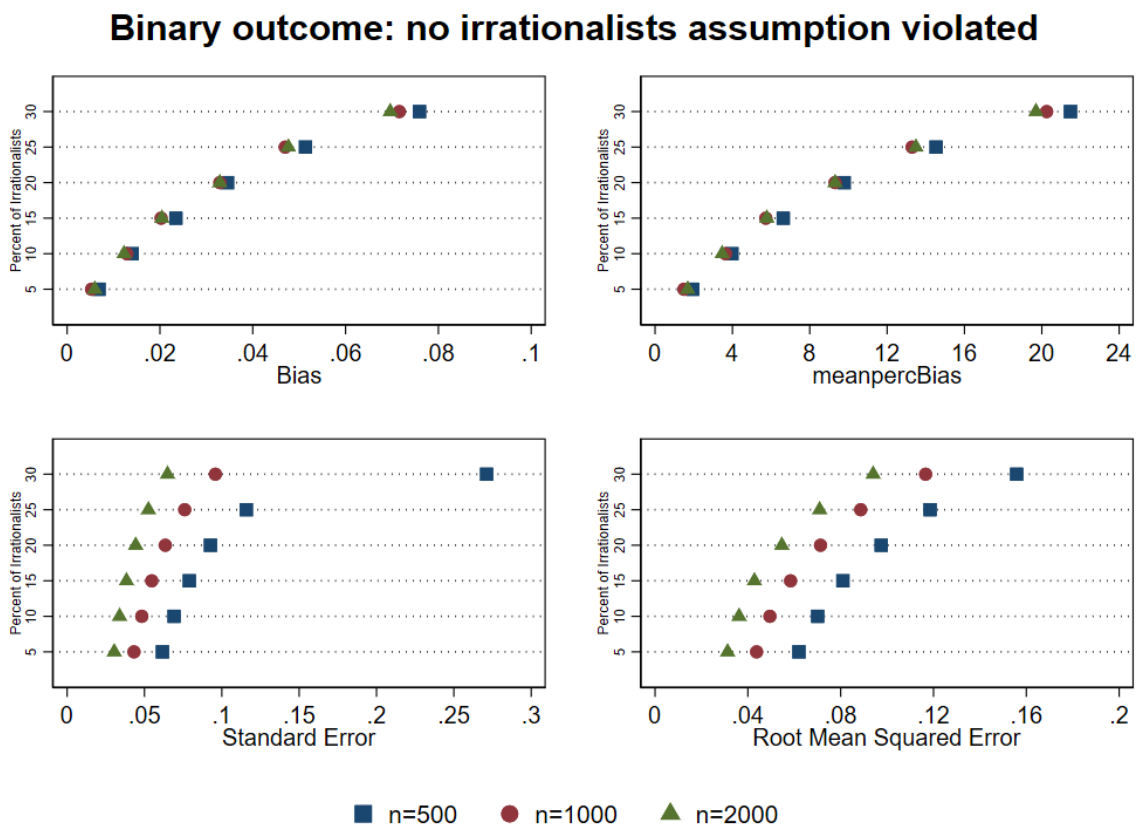

Figure A.3 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of irrationalists for the binary outcome when Assumption II.a, the no irrationalists assumption, is violated (scenario C). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

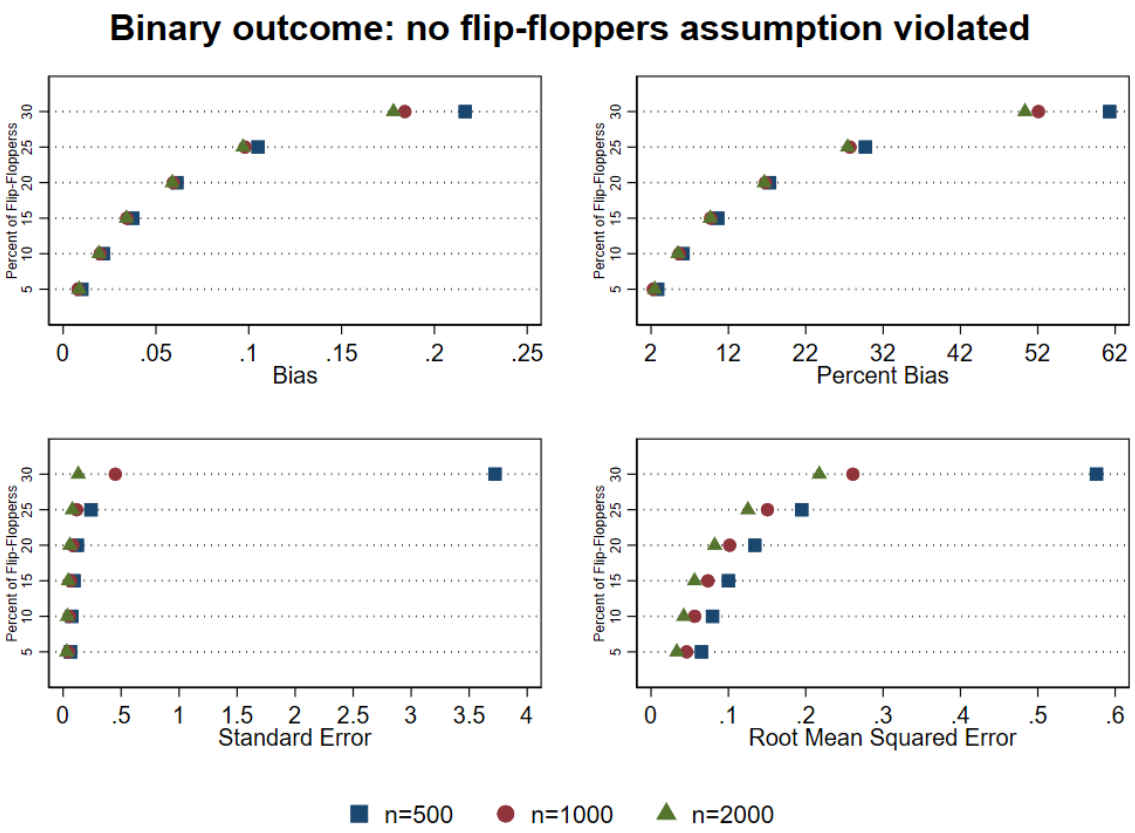

Figure A.4 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of flip-floppers for the binary outcome when Assumption II.b, the no flip-floppers assumption, is violated (scenario D). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

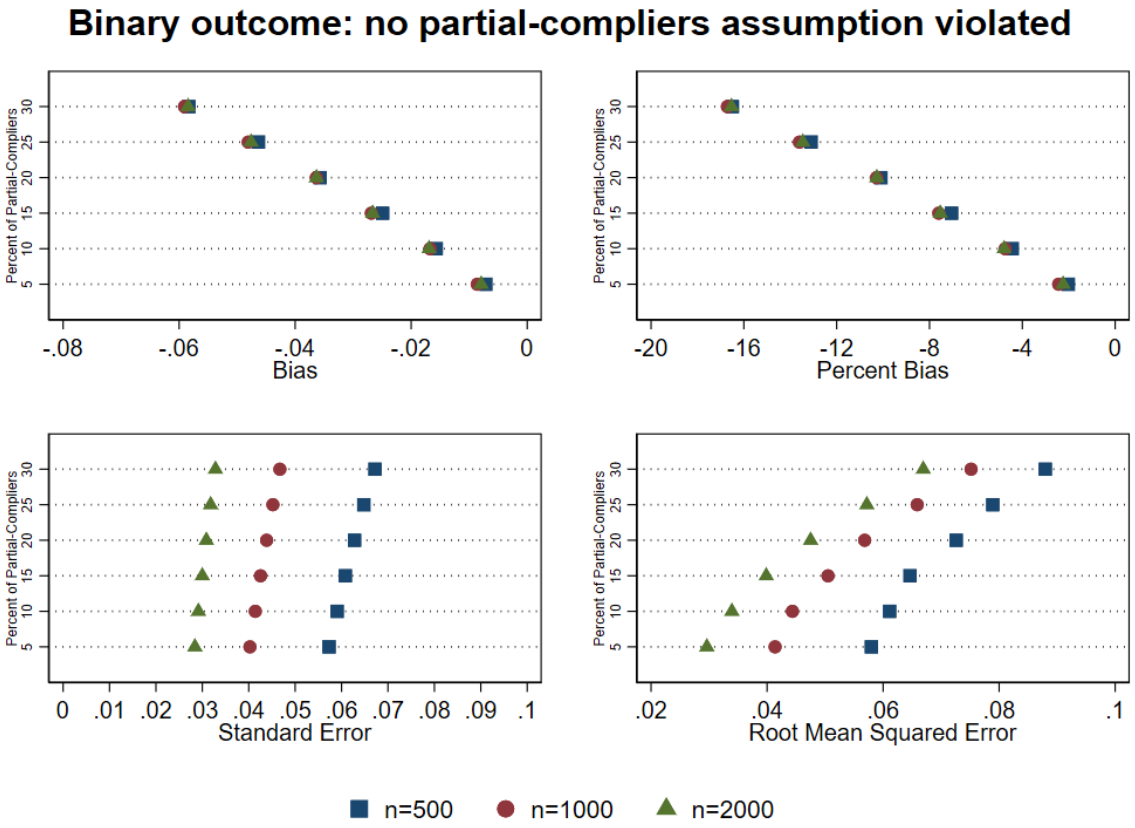

Figure A.5 Sensitivity (upper left: bias; upper right: percent bias; lower left: mean standard error; lower right: root mean squared error) of the nonparametric estimator across proportions of partial-compliers for the continuous outcome when Assumption III, the no partial-compliers assumption, is violated (scenario E). The blue square represents sample size $n = 500$. The red circle represents sample size $n = 1000$. The green triangle represents sample size $n = 2000$.

APPENDIX B

Results of the Binary Outcomes for the Multisite Design Approach

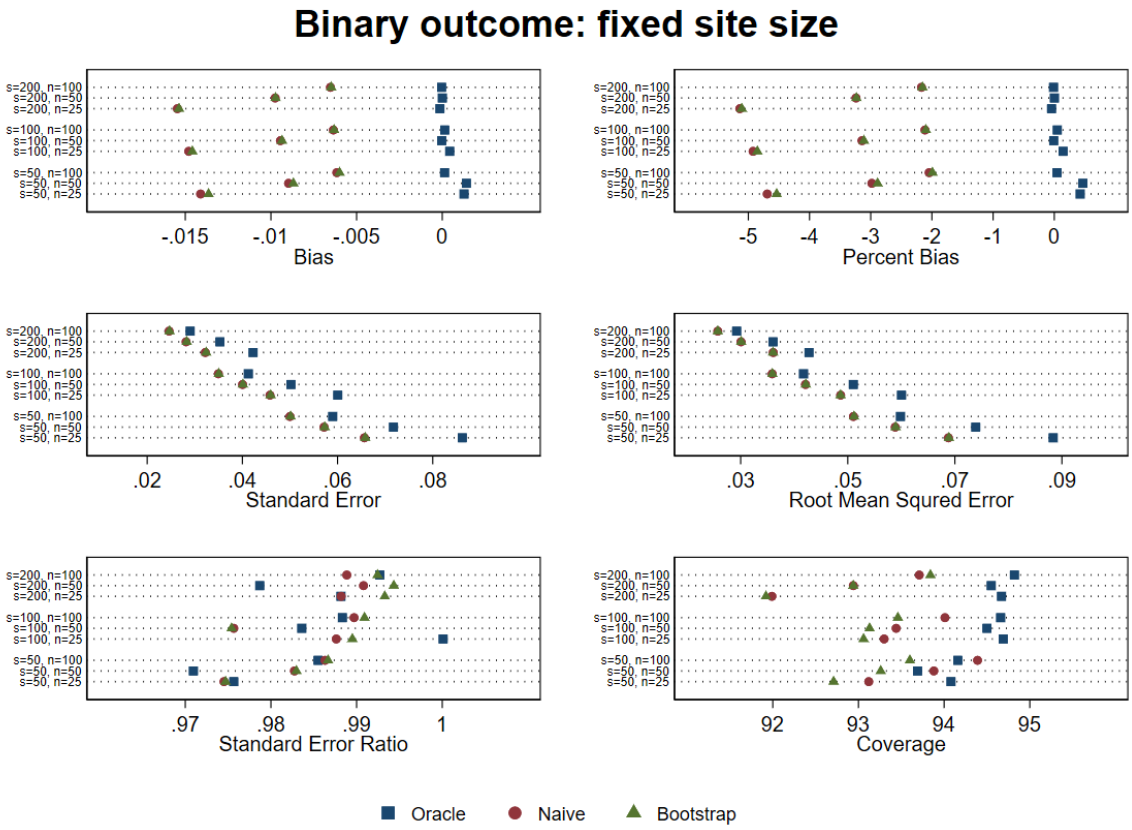

Figure B.1 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the nine scenarios based on number of sites s and site sizes n for the binary outcome when site size is fixed. For example, label "s=50, n=25" indicates that the scenario has 50 sites and 25 individuals in each site. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities.

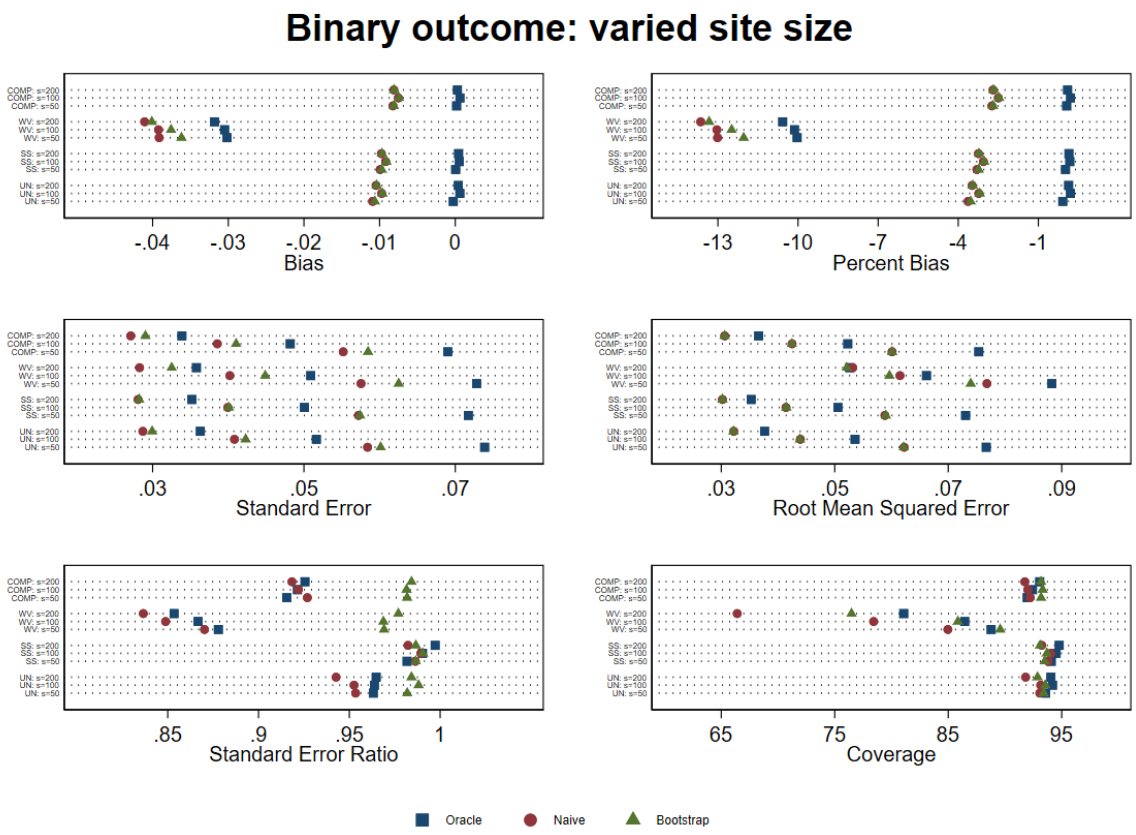

Figure B.2 Performance (top left: bias; top right: percent bias; middle left: mean standard error; middle right: root mean squared error; bottom left: standard error ratio; bottom right: coverage) of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites s for the binary outcome when site size is varied. The blue square represents the oracle estimator which uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The red circle represents the naïve estimator which uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The green triangle represents the bootstrap estimator which is the average of the naïve estimates after resampling the site-level quantities.

Table B.1 Sensitivity of the multisite design estimator across the nine scenarios based on number of sites and site sizes for the binary outcome when site size is fixed if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$.

Measure	# of Sites	Site Size	Oracle			Naïve			Bootstrap		
			C1	C2	C3	C1	C2	C3	C1	C2	C3
Bias (× 100)	50	25	0.13	-8.1	13.2	-1.4	-19.3	-0.64	-1.4	-19.3	-0.60
		50	0.14	-8.1	11.9	-0.90	-16.6	1.8	-0.87	-16.6	1.9
		100	0.01	-8.1	11.0	-0.62	-14.0	4.2	-0.60	-14.0	4.2
	100	25	0.04	-7.9	13.5	-1.5	-19.1	-0.38	-1.5	-19.1	-0.35
		50	0.00	-8.1	12.0	-0.94	-16.7	1.8	-0.94	-16.7	1.8
		100	0.01	-8.2	10.9	-0.64	-14.2	4.1	-0.63	-14.2	4.1
	200	25	-0.01	-8.0	13.3	-1.55	-19.1	-0.51	-1.5	-19.1	-0.49
		50	0.00	-8.0	12.0	-0.97	-16.7	1.8	-0.97	-16.7	1.8
		100	0.00	-8.2	10.8	-0.65	-14.2	4.0	-0.65	-14.2	4.0
Percent Bias (%)	50	25	0.42	-23.6	66.2	-4.7	-56.3	-3.2	-4.5	-56.2	-3.0
		50	0.47	-23.4	60.0	-3.0	-48.3	9.2	-2.9	-48.2	9.3
		100	0.05	-23.4	55.2	-2.0	-40.6	21.0	-2.0	-40.6	21.1
	100	25	0.14	-23.1	67.8	-4.9	-55.8	-1.9	-4.9	-55.7	-1.8
		50	-0.01	-23.4	60.5	-3.1	-48.4	9.1	-3.1	-48.4	9.1
		100	0.05	-23.8	55.0	-2.1	-41.1	20.5	-2.1	-41.1	20.5
	200	25	-0.04	-23.2	67.0	-5.1	-55.8	-2.5	-5.1	-55.7	-2.5
		50	0.00	-23.3	60.2	-3.2	-48.6	8.8	-3.2	-48.5	8.8
		100	-0.01	-23.9	54.5	-2.2	-41.2	20.2	-2.2	-41.2	20.2
Standard Error	50	25	0.09	0.21	0.21	0.07	0.15	0.15	0.07	0.15	0.15
		50	0.07	0.19	0.17	0.06	0.15	0.14	0.06	0.15	0.14
		100	0.06	0.17	0.14	0.05	0.14	0.12	0.05	0.14	0.12
	100	25	0.06	0.15	0.15	0.05	0.11	0.11	0.05	0.10	0.10
		50	0.05	0.13	0.12	0.04	0.10	0.09	0.04	0.10	0.09
		100	0.04	0.12	0.10	0.04	0.10	0.08	0.04	0.10	0.08
	200	25	0.04	0.10	0.10	0.03	0.07	0.07	0.03	0.07	0.07
		50	0.04	0.09	0.08	0.03	0.07	0.07	0.03	0.07	0.07
		100	0.03	0.08	0.07	0.02	0.06	0.06	0.02	0.07	0.06

Table B.1 (cont'd)

Root Mean Square Error	50	25	0.09	0.23	0.25	0.07	0.25	0.15	0.07	0.25	0.15
		50	0.07	0.2	0.21	0.06	0.22	0.14	0.06	0.22	0.14
		100	0.06	0.19	0.18	0.05	0.20	0.12	0.05	0.20	0.13
	100	25	0.06	0.17	0.20	0.05	0.22	0.11	0.05	0.22	0.11
		50	0.05	0.15	0.17	0.04	0.20	0.10	0.04	0.20	0.10
		100	0.04	0.14	0.15	0.04	0.17	0.09	0.04	0.17	0.10
	200	25	0.04	0.13	0.17	0.04	0.20	0.07	0.04	0.20	0.07
		50	0.04	0.12	0.15	0.03	0.18	0.07	0.03	0.18	0.07
		100	0.03	0.12	0.13	0.03	0.16	0.07	0.03	0.16	0.07

Table B.2 Sensitivity of the multisite design estimator across the twelve scenarios based on types of weights (UN: unweighted; SS: weight by site size; WV: weight by within-site variation; COMP: weight by number of compliers) and number of sites for the binary outcome when site size is varied if Assumption V, the zero correlation assumption, is violated. The oracle estimator uses the true principal strata proportions $\pi_{3|k}$, $\pi_{4|k}$ and $\pi_{6|k}$ to fit the multiple linear regression. The naïve estimator uses the estimated principal strata proportions $\hat{\pi}_{3|k}$, $\hat{\pi}_{4|k}$ and $\hat{\pi}_{6|k}$ to fit the multiple linear regression. The bootstrap estimator is the average of the naïve estimates after resampling the site-level quantities. For each of these three estimators, we consider three cases. C1 is the reference case with $\lambda = \gamma = 0$ indicating that Assumption V is satisfied. C2 refers to the case when $\lambda = 2$ and $\gamma = 1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = 1$. C3 refers to the case when $\lambda = 2$ and $\gamma = -1$ indicating that Assumption V is violated and the coefficient of the unobserved confounding is $\gamma = -1$.

Measure	Type of Weight	# of sites	Oracle			Naïve			Bootstrap		
			C1	C2	C3	C1	C2	C3	C1	C2	C3
Bias (× 100)	UN	50	-0.03	-8.0	12.1	-1.1	-17.2	1.3	-1.1	-17.2	1.4
		100	0.06	-8.0	12.2	-0.97	-17.0	1.6	-0.96	-17.0	1.6
		200	0.04	-8.03	12.2	-1.1	-17.1	1.4	-1.0	-17.1	1.5
	SS	50	0.01	-8.0	11.9	-0.99	-16.8	1.7	-0.97	-16.8	1.7
		100	0.05	-8.0	11.9	-0.92	-16.7	1.9	-0.91	-16.7	1.9
		200	0.04	-8.0	12.0	-0.98	-16.7	1.8	-0.97	-16.7	1.8
	WV	50	-0.03	-9.8	9.3	-3.9	-18.5	-0.59	-3.6	-18.4	-0.29
		100	-0.03	-9.9	9.0	-3.9	-18.4	-0.57	-3.8	-18.3	-0.37
		200	-0.03	-10.0	8.8	-4.1	-18.4	-0.86	-4.0	-18.4	-0.73
	COMP	50	0.02	-7.9	11.9	-0.82	-16.8	1.5	-0.81	-16.9	1.5
		100	0.06	-8.0	11.9	-0.75	-16.7	1.8	-0.74	-16.7	1.8
		200	0.03	-8.0	12.0	-0.81	-16.8	1.7	-0.81	-16.8	1.7
Percent Bias (%)	UN	50	-0.09	-23.2	60.7	-3.6	-49.8	6.8	-3.5	-49.7	6.9
		100	0.20	-23.2	61.3	-3.2	-49.4	8.0	-3.2	-49.3	8.0
		200	0.13	-23.3	61.5	-3.5	-49.7	7.2	-3.5	-49.7	7.3
	SS	50	0.02	-23.3	59.9	-3.3	-48.8	8.5	-3.2	-48.8	8.5
		100	0.18	-23.4	60.0	-3.1	-48.3	9.4	-3.0	-48.4	9.4
		200	0.15	-23.3	60.5	-3.2	-48.5	9.1	-3.2	-48.5	9.1
	WV	50	-10.0	-28.5	46.9	-13.0	-53.7	-3.0	-12.0	-53.3	-1.4
		100	-10.1	-28.7	45.3	-13.0	-53.4	-2.9	-12.5	-53.2	-1.9
		200	-10.6	-28.9	44.5	-13.6	-53.5	-4.3	-13.3	-53.4	-3.7
	COMP	50	0.06	-22.9	60.0	-2.7	-48.8	7.8	-2.7	-48.9	7.7
		100	0.21	-23.3	60.0	-2.5	-48.6	8.9	-2.5	-48.6	8.9
		200	0.10	-23.3	60.2	-2.7	-48.7	8.5	-2.7	-48.7	8.5

Table B.2 (cont'd)

Standard Error	UN	50	0.07	0.19	0.18	0.06	0.15	0.14	0.06	0.15	0.14
		100	0.05	0.13	0.12	0.04	0.10	0.10	0.04	0.10	0.10
		200	0.04	0.09	0.09	0.03	0.07	0.07	0.03	0.07	0.07
	SS	50	0.07	0.19	0.17	0.06	0.15	0.14	0.06	0.15	0.14
		100	0.05	0.13	0.12	0.04	0.10	0.09	0.04	0.10	0.09
		200	0.04	0.09	0.08	0.03	0.07	0.07	0.03	0.07	0.07
	WV	50	0.07	0.19	0.17	0.06	0.15	0.13	0.06	0.16	0.14
		100	0.05	0.13	0.12	0.04	0.10	0.09	0.04	0.12	0.10
		200	0.04	0.09	0.08	0.03	0.07	0.06	0.03	0.08	0.07
	COMP	50	0.07	0.20	0.18	0.06	0.15	0.14	0.06	0.15	0.14
		100	0.05	0.14	0.12	0.04	0.10	0.10	0.04	0.11	0.10
		200	0.03	0.10	0.09	0.03	0.07	0.07	0.03	0.07	0.07
Root Mean Squared Error	UN	50	0.08	0.21	0.22	0.06	0.23	0.15	0.06	0.23	0.15
		100	0.05	0.16	0.18	0.04	0.20	0.10	0.04	0.20	0.10
		200	0.04	0.12	0.15	0.03	0.19	0.07	0.03	0.19	0.07
	SS	50	0.07	0.21	0.21	0.06	0.22	0.14	0.06	0.22	0.14
		100	0.05	0.15	0.17	0.04	0.20	0.10	0.04	0.20	0.10
		200	0.04	0.12	0.15	0.03	0.18	0.07	0.03	0.18	0.07
	WV	50	0.09	0.25	0.21	0.08	0.25	0.14	0.07	0.25	0.14
		100	0.07	0.19	0.16	0.06	0.22	0.10	0.06	0.22	0.10
		200	0.05	0.15	0.13	0.05	0.20	0.07	0.05	0.20	0.07
	COMP	50	0.08	0.22	0.22	0.06	0.23	0.14	0.06	0.23	0.14
		100	0.05	0.16	0.17	0.04	0.20	0.10	0.04	0.20	0.10
		200	0.04	0.13	0.15	0.03	0.18	0.07	0.03	0.18	0.07

BIBLIOGRAPHY

BIBLIOGRAPHY

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). *Identification of Causal Effects Using Instrumental Variables*. 13.

Balke, A., & Pearl, J. (1994). Counterfactual Probabilities: Computational Methods, Bounds and Applications. In *Uncertainty Proceedings 1994* (pp. 46–54). Elsevier.

Balke, A., & Pearl, J. (1997). Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176.

Bech, P., Fava, M., Trivedi, M. H., Wisniewski, S. R., & Rush, A. J. (2012). Outcomes on the pharmacopsychometric triangle in bupropion-SR vs. buspirone augmentation of citalopram in the STAR*D trial: Outcomes on the pharmacopsychometric triangle. *Acta Psychiatrica Scandinavica*, *125*(4), 342–348.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842.

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F., Kupfer, D. J., & for the STAR∗D Investigators Group. (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR∗D) study. *Psychiatric Clinics of North America*, *26*(2), 457–494.

Frangakis, C. E. (2002). Clustered Encouragement Designs with Individual Noncompliance: Bayesian Inference with Randomization, and Application to Advance Directive Forms. *Biostatistics*, *3*(2), 147–164.

Frangakis, C. E., & Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, *58*(1), 21–29.

Frank, E., Karp, J. F., & Rush, A. J. (1996). Efficacy of treatments for major depression. *Psychopharmacology Bulletin*, 29(4), 457-475.

Gelenberg, A. J., Freeman, M. P., Markowitz, J. C., Rosenbaum, J. F., Thase, M. E., Trivedi, M. H., Rhoads, R. S. V., Reus, V. I., DePaulo, J. R., Fawcett, J. A., Schneck, C. D., & Silbersweig, D. A. (2010). *WORK GROUP ON MAJOR DEPRESSIVE DISORDER*. 152.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). The Economic Burden of Adults With Major Depressive Disorder in the United States (2005 and 2010). *The Journal of Clinical Psychiatry*, *76*(02), 155–162.

Hernán, M. A., & Robins, J. M. (2019). *Causal Inference*.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*(1), 69–88.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*, 467–475.

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6), 1845-1857.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327. https://doi.org/10.1214/aos/1034276631

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

ICH E9 working group. (2020). ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … Murray, C. J. L. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858.

Jiang, Z., Ding, P., & Geng, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(4), 829–848.

Jo, B. (2002). Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. *Journal of Educational and Behavioral Statistics*, *27*(4), 385–409.

Little, R. J., Long, Q., & Lin, X. (2009). A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. *Biometrics*, *65*(2), 640–649.

Little, R. J., Long, Q., & Lin, X. (2009). A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. *Biometrics*, *65*(2), 640–649.

Long, Q., Little, R. J. A., & Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: A Bayesian framework. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(3), 513–531.

Mehrotra, D. V. (2019). A note on the draft International Council for Harmonisation guidance on estimands and sensitivity analysis. *Clinical Trials*, *16*(4), 339–344.

Mehrotra, D. V., Liu, F., & Permutt, T. (2017). Missing data in clinical trials: Control-based mean imputation and sensitivity analysis. *Pharmaceutical Statistics*, *16*(5), 378–392.

Miratrix, L., Furey, J., Feller, A., Grindal, T., & Page, L. C. (2018). Bounding, An Accessible Method for Estimating Principal Causal Effects, Examined and Explained. *Journal of Research on Educational Effectiveness*, *11*(1), 133–162.

National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press.

Pearl, J. (2011). Principal Stratification—A Goal or a Tool? *The International Journal of Biostatistics*, *7*(1), 1–13.

Pigott, H. E. (2015). The STAR*D Trial: It is Time to Reexamine the Clinical Beliefs That Guide the Treatment of Major Depression. *The Canadian Journal of Psychiatry*, *60*(1), 9–13.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. *American Journal of Evaluation*, *36*(4), 475–499.

Reardon, S. F., & Raudenbush, S. W. (2013). Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects? *Sociological Methods & Research*, *42*(2), 143–163.

Reardon, S. F., Unlu, F., Zhu, P., & Bloom, H. S. (2014). Bias and Bias Correction in Multisite Instrumental Variables Analysis of Heterogeneous Mediator Effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 53–86.

Roy, J., Hogan, J. W., & Marcus, B. H. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, *9*(2), 277–289.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., Kupfer, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J. W., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B.

D., Ritz, L., … for the STAR*D Investigators Group. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Controlled Clinical Trials*, *25*(1), 119–142.

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Niederehe, G., Thase, M. E., Lavori, P. W., Lebowitz, B. D., McGrath, P. J., Rosenbaum, J. F., & Sackeim, H. A. (2006). Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several Treatment Steps: A STAR*D Report. *Am J Psychiatry*, 13.

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Stewart, J. W., Nierenberg, A. A., Thase, M. E., Ritz, L., Biggs, M. M., Warden, D., Luther, J. F., Shores-Wilson, K., Niederehe, G., & Fava, M. (2006). Bupropion-SR, Sertraline, or Venlafaxine-XR after Failure of SSRIs for Depression. *New England Journal of Medicine*, *354*(12), 1231–1242.

Shrier, I., Steele, R. J., Verhagen, E., Herbert, R., Riddell, C. A., & Kaufman, J. S. (2014). Beyond intention to treat: What is the right question? *Clinical Trials: Journal of the Society for Clinical Trials*, *11*(1), 28–37.

Sinyor, M., Schaffer, A., & Levitt, A. (2010). The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Trial: A Review. *The Canadian Journal of Psychiatry*, *55*(3), 126–135.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, *5*(4), 465–472.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., & Richardson, T. S. (2018). Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes. *Journal of the American Statistical Association*, *113*(522), 933–947.

Trivedi, M. H., Thase, M. E., Ritz, L., Biggs, M. M., & Rush, A. J. (2006). Medication Augmentation after the Failure of SSRIs for Depression. *The New England Journal of Medicine*, 10.

VanderWeele, T. J. (2011). Principal Stratification—Uses and Limitations. *The International Journal of Biostatistics*, *7*(1), 1–14.

World Health Organization. *(2017). Depression and other common mental disorders: global health estimates* (No. WHO/MSD/MER/2017.2). World Health Organization.

Yuan, L.-H., Feller, A., & Miratrix, L. W. (2018). Identifying and Estimating Principal Causal Effects in Multi-site Trials. *ArXiv:1803.06048 [Stat]*.