# Methods for Causal Inference in Randomized Trials with Multiple Versions of Control and Noncompliance, with an Application to Behavioral Intervention Trials

Scott Coggeshall

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Xiao-Hua (Andrew) Zhou, Chair

Thomas Richardson

Yanqin Fan

Program Authorized to Offer Degree:
Biostatistics

University of Washington

**Abstract**

Methods for Causal Inference in Randomized Trials with Multiple Versions of Control and Noncompliance, with an Application to Behavioral Intervention Trials

Scott Coggeshall

Chair of the Supervisory Committee:
Professor Xiao-Hua (Andrew) Zhou
Biostatistics

Behavioral therapies are a class of interventions with a wide array of applications. Because of the complicated nature of these interventions, however, conducting randomized controlled trials of these interventions poses unique challenges compared to the classical blinded, placebo-controlled RCT. The primary issue is that RCTs of behavioral interventions often use treatment-as-usual (TAU) control groups, due to the lack of a feasible "placebo" equivalent to the active intervention. As a result, control groups in these trials are typically heterogeneous with respect to the form of treatment received, making causal inference under the standard assumption of "no multiple versions of treatment" no longer applicable. In this dissertation, we develop frameworks for causal inference in single-site and multi-site RCTs with multiple versions of control due to the use of a TAU control group. We define causal estimands of interest based on a principle stratification approach. We show that these causal estimands are only partially identified with data from a single-site RCT, but can be identified under certain assumptions with data from a multi-site RCT. We then propose methods for performing inference for these causal estimands, either through bounding (in the case of partial identifiability) or point estimation (in the case of identifiability). Finally, we apply these methods to an RCT of a behavioral therapy intervention for children with autism. Additional work in this dissertation includes an examination of identifiability issues with methods for causal inference in RCTs with partial compliance, a tutorial for a Bayesian approach to binary non-compliance in RCTs, and a systematic review of behavioral interventions for children with autism.

# TABLE OF CONTENTS

Page

List of Figures . . . . . iv

List of Tables . . . . . vii

Chapter 0: Introduction . . . . . 1

Chapter 1: Causal Inference for RCTs with Multiple Versions of Control . . 3

1.1 Introduction . . . . . 3

1.2 Notation . . . . . 6

1.3 Assumptions . . . . . 8

1.4 Effect of Control Group Heterogeneity on Generalizability of the $ITT$ Effect . . . . . 11

1.5 Identifiability of $\Delta_{01}$ and $\Delta_{21}$ . . . . . 14

1.6 Linear Programming Bounds for $\Delta_{01}$ and $\Delta_{21}$ . . . . . 18

1.7 Checking the Model Assumptions . . . . . 29

1.8 Simulations . . . . . 31

1.9 Application to Early Start Denver Model RCT . . . . . 35

1.10 Discussion . . . . . 38

Chapter 2: Causal Inference in RCTs with Multiple Versions of Control with Data from Multiple Populations . . . . . 43

2.1 Introduction . . . . . 43

2.2 Notation . . . . . 45

2.3 Assumptions . . . . . 46

2.4 Estimation and Inference for Observed-data Quantities . . . . . 50

2.5 Identifiability with Data from Two Populations under the Exclusion Restriction . . . . . 52

2.6 Identifiability with Data from $k > 2$ Populations under the Exclusion Restriction Assumption . . . . . 54
2.7 Simulations for $\hat{\mathbf{\Delta}}$ and $\hat{\mathbf{\Delta}}_{LS}$ . . . . . 56
2.8 Re-analysis of the ESDM RCT data . . . . . 62
2.9 Identifiability with Data from Two Populations without the Exclusion Restriction . . . . . 64
2.10 Identifiability with Data from $K \geq 2$ Populations without the Exclusion Restriction . . . . . 66
2.11 Simulations for Multi-site Bounds on $\Delta_{01}$ and $\Delta_{21}$ without the Exclusion Restriction . . . . . 72
2.12 Comparable Strata Effects Assumption and Treatment Effect Heterogeneity . . . . . 77
2.13 Discussion . . . . . 79

Chapter 3: Causal Inference with Partial Compliance to Treatment . . . . 83
3.1 Introduction . . . . . 83
3.2 Notation and Assumptions for Counterfactuals . . . . . 84
3.3 Identifability Concerns in Partial Compliance Models . . . . . 85
3.4 Likelihood-Based Approaches . . . . . 86
3.5 Simulations Demonstrating Inconsistency of Maximum likelihood-based Partial Compliance Methods . . . . . 92
3.6 Simulations Demonstrating Bias under Misspecification of Correlation between Potential Compliance Behaviors . . . . . 96
3.7 Discussion . . . . . 100

Chapter 4: Analyzing Randomized Trials with Non-Compliance Using the `noncomplyR` R Package . . . . . 102
4.1 Preface . . . . . 102
4.2 Introduction . . . . . 103
4.3 Causal Inference, Potential Outcomes and Adherence . . . . . 105
4.4 Implementation of the Data Augmentation Algorithm with the `noncomplyR` Package . . . . . 114
4.5 Discussion . . . . . 124

Chapter 5: A Systematic Review of Effectiveness Studies of Behavioral Interventions for Children with ASD . . . . . 126
5.1 Introduction . . . . . 126
5.2 Methods . . . . . 127
5.3 Results . . . . . 129
5.4 Forest Plots and Meta-Analyses . . . . . 132
5.5 Discussion . . . . . 135
5.6 Limitations . . . . . 137
5.7 Conclusion . . . . . 137

Chapter 6: Discussion . . . . . 139

Appendix A: . . . . . 151
A.1 Proofs . . . . . 151
A.2 Functions Related to Large Sample Distributions of Bounds Estimators 158

Appendix B: . . . . . 161

Appendix C: . . . . . 165

Appendix D: . . . . . 167

# LIST OF FIGURES


Figure Number Page

1.2 Visualization of the linear constraints on $p_{01\cdot1}$ and $p_{21\cdot1}$ under Assumptions 1-5. The panels show the examples of possible combinations of identifiability. In Panel A, $p_{01\cdot1}$ is partially identified while $p_{21\cdot1}$ is fully unidentified. In Panel B, $p_{21\cdot1}$ is partially identified while $p_{01\cdot1}$ is fully unidentified. In Panel C, both $p_{01\cdot1}$ and $p_{21\cdot1}$ are fully unidentified. In Panel D, both $p_{01\cdot1}$ and $p_{21\cdot1}$ are asymmetrically partially identified. 24

1.3 Trace plots showing consistency of estimates of $\Delta_{01}^{U,ER}$ and $\Delta_{01}^{L,ER}$ under several identifiability scenarios for the causal parameter $p_{01\cdot1}$. In the upper lefthand panel, the causal estimand $\Delta_{01}$ is asymmetrically partially identified with uninformative bounds. In the upper righthand panel, $\Delta_{01}$ is partially identified, but the bounds are uninformative. In the lower lefthand panel, $p_{01\cdot1}$ is partially identified, with informative lower and upper bounds. In each panel, the red (solid) horizontal lines are the true values for $\Delta_{01}^{U,ER}$ and $\Delta_{01}^{L,ER}$ . . . . . 33

1.5 Estimated coverage probabilities for 95% confidence intervals as a function of the parameter $\omega_{01}$. Panel A shows coverages for the bootstrap confidence intervals for the linear programming-based bounds. Panel B shows coverages for the confidence intervals based on asymptotic distributions. In both panels, the figure on the left shows estimated coverages at $n = 1000$, while the figure on the right shows estimated coverages at $n = 10000$. . . . . 34

1.6 Scatterplot of individual versus group average treatment hours at one year post-randomization. The ESDM treatment group is shown in red, while the TAU control group is shown in blue. . . . . 37

2.1 Trace plots showing the convergence of the estimator $\hat{\Delta}_{LS,01}$ to the true value of the general multi-site estimator $\Delta_{01}$ (indicated by the red solid line) as the sample size increases. The top panel shows results from the scenario where $\omega_{01}$, the proportion of individuals belonging to the $D(0) = 0, D(1) = 1)$ principal stratum, was set to a relatively low value in all three sites. The middle value shows results from the scenario where $\omega_{01}$ was set to a value around 0.5 in all three sites. The bottom panel shows results from the scenario where $\omega_{01}$ was set to a relatively high value in all three sites. . . . . 59

2.2 Trace plots showing the convergence of the two-site specific estimator $\hat{\Delta}_{01}$ to the true value of $\Delta_{01}$ (indicated by the red solid line) as the sample size increases. The top panel shows results from the scenario where $\omega_{01}$, the proportion of individuals belonging to the $D(0) = 0, D(1) = 1)$ principal stratum, was set to a relatively low value in all three sites. The middle value shows results from the scenario where $\omega_{01}$ was set to a value around 0.5 in all three sites. The bottom panel shows results from the scenario where $\omega_{01}$ was set to a relatively high value in all three sites. . . . . 60

2.3 Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is relatively large within each site. . . . . 74

2.4 Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is in the medium range within each site. . . . . 75

2.5 Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is relatively small within each site. . . . . 76

3.1 Trace plots showing linear regression parameter estimates from the two-stage maximum-likelihood approach and the weighted EM algorithm approach as a function of increasing sample size. In each panel, the blue line shows the two-stage estimates, the black line shows the weighted EM estimates, and the red line shows the true underlying value for the regression parameter. . . . . 94

3.2 Array of trace plots showing estimates of selected points on the $PCE$ surface from a linear outcome model as a function of increasing sample size. Each cell in the $3x3$ array shows estimates for $PCE(D(0), D(1))$, e.g. the upper left-hand cell shows estimates for $PCE(0.6, 0.6)$. In each cell, the blue line shows estimates from the two-stage maximum-likelihood approach, the black line shows estimates from the weighted EM algorithm, and the red line shows the true value for $PCE(D(0), D(1))$. In each cell, the x-axis ranges from $n = 1000$ to $n = 100000$. . . . . 95

3.3 Trace plots showing logistic regression parameter estimates from the two-stage maximum-likelihood approach and the weighted EM algorithm approach as a function of increasing sample size. In each panel, the blue line shows the two-stage estimates, the black line shows the weighted EM estimates, and the red line shows the true underlying value for the regression parameter. . . . . 97

3.4 Array of trace plots showing estimates of selected points on the $PCE$ surface from a logistic outcome model as a function of increasing sample size. Each cell in the $3x3$ array shows estimates for $PCE(D(0), D(1))$, e.g. the upper left-hand cell shows estimates for $PCE(0.6, 0.6)$. In each cell, the blue line shows estimates from the two-stage maximum-likelihood approach, the black line shows estimates from the weighted EM algorithm, and the red line shows the true value for $PCE(D(0), D(1))$. In each cell, the x-axis ranges from $n = 1000$ to $n = 100000$. . . . . 98

3.5 Mean values for the regression parameter estimates from the weighted EM algorithm as a function of the assumed value for $\rho$ used in fitting the model. In each panel, the red line shows the true regression parameter value used in generating the data. . . . . 99

4.1 Posterior distributions of the Complier Average Causal Effect, both without the exclusion restriction assumption (left panel) and with the exclusion restriction assumption (right panel). . . . . 119

4.2 Left panel: Posterior distribution of the CACE from the ESDM dataset under an uninformative prior. Right panel: Posterior distribution of the CACE from the ESDM dataset under a user-specified informative prior. . . . . 123

5.1 Forest plot of estimates for the Vineland Adaptive Behavior Scale outcome. . . . . 133

5.2 Forest plot of estimates for the cognitive development outcome. . . . 134

5.3 Forest plot of estimates for joint attention outcome. . . . . 134

# LIST OF TABLES


Table Number Page

1.1 Principal strata and their associated probability parameters. Dashes indicate the principal strata that are ruled out by the self-motivated treatment assumption. . . . . 11

1.2 Observed-data distributions implied by causal parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The agreement between the two observed-data distributions despite different underlying values for $\Delta_{01}$ and $\Delta_{21}$ demonstrates the identifiability issue. . . . . 17

1.3 Point and interval estimates for the $ITT$ effect and $\Delta_{01}$ at 1 year post-randomization. . . . . 38

2.1 Underlying causal model parameter values for the three sites under the three different data generating scenarios. . . . . 58

2.2 Estimated coverage of 95% confidence intervals for $\Delta_{01}$ with data from multiple populations under Assumptions WS-ER. . . . . 61

2.3 Observed-data distributions implied by the underlying causal parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$. The equality of the two observed-data distributions despite the different underlying values for $\Delta_{01}$ and $\Delta_{21}$ demonstrates the identifiability issue for these causal estimands without the exclusion restriction assumption. . . . . 65

2.4 True values for the causal parameter vectors for a three-site randomized trial with multiple versions of control. The values of the causal parameter vectors were set so that Assumptions WS-NER hold. . . . 72

2.5 Estimated empirical bias of $\widehat{\Delta}_{01}^{L,MS}$ and $\widehat{\Delta}_{01}^{U,MS}$ at different sample sizes and for different values of $\omega_{01}$ . . . . . 73

4.1 Possible compliance types in the binary non-compliance framework. . 109

5.1 Characteristics of the review studies related to methodological quality. 135

C.1 Complete lists of parameters and hyperparameters for the binary and Normal outcome models under four sets of assumptions. ER = Exclusion Restriction assumption, SA = Strong Access Monotonicity assumption . . . . . 166

D.1 Characteristics of the review studies. . . . . 168

# ACKNOWLEDGMENTS

I would like to thank my advisor, Andrew Zhou, for his guidance and insights. I thank the members of my committee for their participation and valuable contributions to this work. I thank Annette Estes, Jeffrey Munson, the UW Autism Center, and the Center on Human Development and Disabilities for their support throughout my time here at UW as well as for the use of the data from the ESDM RCT. I thank the UW Department of Biostatistics for giving me the opportunity to study here. Finally, I thank my fellow students in the Department of Biostatistics.

# DEDICATION

To my parents, Jack and Donna, for their endless support and encouragement. And to my dog Sheldon, who was my companion for much of the writing of this thesis.

# Chapter 0

# INTRODUCTION

This dissertation represents my contributions to what I believe are three key roles for statisticians: researchers, educators, and scientists. First, as statisticians we are primarily responsible for the continued development of the field of statistics. This involves both the development of novel methodologies to address gaps in the literature, as well as the investigation and critique of existing methods. In Chapters 1 and 2, I study in-depth the case of RCTs that violate the standard assumption in causal inference of "no multiple versions of treatment/control" due to the use of a treatment-as-usual (TAU) control group in place of a placebo-based control group. I propose a novel framework for performing causal inference for RCTs with multiple versions of control. Using this framework, I define two causal estimands of interest and examine their identifiability in both single-site and multi-site trials. I then develop methodology for performing inference about these estimands, either through bounds (when they are unidentified) or through point estimation (when they are identified). The motivation for this work comes from an RCT of the Early Start Denver Model, a behavioral intervention for children with autism [Dawson et al., 2010].

In Chapter 3, I critically examine some previously published methods for compliance-based causal analyses when compliance is treated as a continuous variable. I demonstrate that these methods produce unreliable estimates of their target for inference, and hence are not suitable for use. I then propose and investigate an alternative method, but show that identifiability issues inherent in models for partial compliance in RCTs prevent this method from being useful in practice.

In addition to our role as developers of statistical methodology, statisticians also have an important role to play as educators. The demand for statistical analyses far exceeds the supply of trained statisticians, meaning that many statistical analyses are performed by researchers without extensive statistical or mathematical backgrounds.

Statisticians often lament the statistical errors and misconceptions that appear in the scientific literature. At the same time, efforts at improving statistical education seem primarily focused on reforms in the classroom or textbooks, effectively placing the burden of improving statistical education on the comparatively few statisticians with the proclivity for teaching and writing textbooks. It is my view that every statistician can serve a valuable role as a statistical educator through the writing of clear, tutorial-style articles for more advanced methodologies. Chapter 4 presents the results of a tutorial-style paper aimed at clearly explaining to non-statisticians how an analysis of data from a randomized controlled trial with non-compliance can be conducted. In conjunction with this tutorial, I have developed the R package noncomplyR, which allows for simple implementation of special cases of the data augmentation algorithm for fitting binary non-compliance models to data from RCTs.

Finally, statisticians also have a role to play as scientists and subject-matter specialists in our own right. By having a deep understanding of a particular field of scientific research, we improve our applied statistical analyses and have a better opportunity to identify areas where new statistical methodology is needed. Equally important, by showing that we can come to the table as fellow scientists, we gain credibility with our scientific collaborators and thus can make a greater impact. Since my dissertation work has been heavily influenced by issues related to conducting RCTs of behavioral interventions for children with autism, in Chapter 5 I present a systematic review of the literature related to these interventions.

In Chapter 6 I summarize the main points of this dissertation and discuss some possibilities for future research.

# Chapter 1

# CAUSAL INFERENCE FOR RCTS WITH MULTIPLE VERSIONS OF CONTROL

## *1.1 Introduction*

A key consideration in the design, implementation, and interpretation of a randomized controlled trial (RCT) is the choice of the control group to which the group receiving the intervention of interest will be compared. In RCTs of pharmacological interventions, it is common to use a placebo, an inert version of the intervention, as a control condition [Millum and Grady, 2013]. Using a placebo for the control condition offers several advantages. First, it aids in blinding the investigators and participants to treatment assignment, which reduces the potential for bias [Juni et al., 2001]. Second, it can typically be produced with a great deal of uniformity. This uniformity of the control condition helps to ensure that the individuals assigned to the control group receive the same version of that control, an important requirement for conducting standard causal inference using the potential outcomes framework [Schwartz et al., 2012]. For many types of intervention, however, a placebo-type version of the intervention is not available to serve as the control condition. An important category of interventions that lack a placebo-type control are behavioral therapy interventions. These interventions involve structured interpersonal interactions in sessions that may last for hours at a time. A placebo-type equivalent to this type of intervention is thus difficult to develop or implement. As a result, RCTs for these interventions employ alternative forms of control groups that are more likely to suffer from control group heterogeneity. Control group heterogeneity in this context refers to the situation where individuals assigned to the control group receive one of several different versions of the control condition post-randomization. In this paper, we develop a novel framework for analyzing data from randomized trials that exhibit a particular pattern of heterogeneity in the control group that can arise when a study uses assignment to treatment-as-usual (TAU) as the control condition.

Our framework is motivated by a specific pattern of control group heterogeneity often seen in behavioral therapy interventions for children with autism spectrum disorder (ASD). These interventions typically involve children receiving multiple hours per week of one-on-one therapy over a span of multiple months. Implementing a placebo-equivalent to these types of interventions is therefore infeasible. As an alternative, researchers often use a TAU (sometimes called community-based care) control condition [Dawson et al., 2010] [Rogers et al., 2012]. The parents of children assigned to the control group receive information about and referrals to treatment options available in the surrounding community. The actual care received by children in the control group is therefore determined by the children's caretakers post-randomization and may be rather heterogeneous. Some children assigned to the control condition may receive no treatment for their ASD. Some may receive a type of behavior therapy treatment that is identical or highly similar to the intervention being studied. Finally, some children in the control arm may receive a form of treatment that is distinct from the intervention being studied. Each of these scenarios represents a distinct version of the TAU control condition. Furthermore, we would expect that a child assigned to TAU who receives no treatment for his or her ASD would likely have a different outcome than if that same child had been assigned to TAU but had received some form of treatment. Using the terminology of the Rubin causal model [Rubin, 1974], the potential outcomes for the children in the control group are not determined solely by their assignment to the control group. As a result, the heterogeneous character of the community-based control can be viewed as a violation of the "no-multiple-versions-of-treatment" component of the Stable Unit Treatment Value Assumption (SUTVA) [Rubin, 1980]. In this chapter, we introduce a framework for causal inference for binary outcomes when the SUTVA is violated due to the type of control group heterogeneity just described, in which individuals in the control group can be classified as having received either no treatment, treatment very similar to the intervention being studied, or treatment distinct from the intervention being studied. We demonstrate how certain causal estimands of interest can be bounded based on the observed data and construct consistent estimators for these bounds.

The SUTVA contains two key components: non-interference and no multiple versions of treatment. Non-interference states that an individual's potential outcomes are not affected by the treatment status of the other individuals in the population.

Violations of this assumption, and methods for addressing it, have received attention in the literature on statistical methods for assessing the efficacy and effectiveness of vaccines [Hudgens and Halloran, 2008]. The "no multiple versions of treatment" component of the SUTVA states that there is only a single version of the treatment and a single version of the control, so that formulating the potential outcomes in terms of the potential outcome under treatment and the potential outcome under control is well-defined. When the "no multiple versions of treatment" component of the SUTVA is violated, this articulation of the potential outcomes is no longer applicable. Hence, much of the initial focus on addressing violations of the SUTVA has dealt with the related question of consistency, or how to connect the potential outcomes to the actual observed outcomes in the data [Pearl, 2010] [VanderWeele, 2009] [Cole and Frangakis, 2009]. More recently, general frameworks for causal inference with multiple versions of treatment have been proposed by Hernán and VanderWeele [2011], VanderWeele and Hernan [2013], with Petersen [2011] discussing how treatments with multiple versions can be understood in terms of causal graphs. Hasegawa et al. [2017] consider randomization-based inference in the case of a treatment or control with two versions. These previous works have primarily focused on estimation of marginal causal effects in situations where observational data are collected on an outcome of interest, an indicator for the treatment assignment, and additional variables that may or may not include the version of treatment received. In VanderWeele and Hernan [2013], the authors focus on defining and identifying several types of marginal causal effects. Their approach to identification and estimation is based on conditioning on and subsequently averaging over relevant confounding variables. They delineate identification and estimation of these marginal effects in several scenarios, including when information on the version of treatment is observed or unobserved. Crucial to their approach are the assumptions that data are available on all relevant confounders, and that the confounders can be reliably categorized into two groups: confounders of the treatment-outcome relationship and confounders of the treatment-version relationship. While their approach could in theory be used to account for the control group heterogeneity scenario we consider, the reliance on confounder adjustment would fail to take advantage of the randomized nature of the treatment assignment, essentially treating the randomized trial as if it were an observational study. The assumption that all relevant confounders have been both measured and reliably identified as either

a treatment-version confounder or version-outcome confounder is strong and unlikely to hold in practice.

A novel contribution of our approach is to work within a principal stratification framework [Frangakis and Rubin, 2002a] and exploit the randomized treatment assignment as an instrumental variable [Angrist et al., 1996], thereby avoiding the need to consider confounding variables when accounting for the version of treatment received. As a result, our method only requires collecting data on the outcome, the assigned treatment, and the version of treatment received. Another contribution of our framework is that it aids researchers in understanding how the results from an RCT with control group heterogeneity may generalize to other populations of interest. Hernán and VanderWeele [2011] discuss how the presence of multiple versions of treatment affects the generalizability or transportability of causal effects. While the authors do mention the case where the data come from a randomized experiment, they deemphasize this scenario in favor of focusing on the case of observational data. The reason for this is that certain causal estimands of interest, most notably the Intent-to-Treat ($ITT$) effect, can be shown to still be identifiable and estimable from a randomized experiment, even when multiple versions of treatment are present. The implication is that the presence of multiple versions of treatment can be safely ignored in the setting of a randomized trial by focusing on the $ITT$ effect. However, we demonstrate in the following that the $ITT$ effect estimated by an RCT with multiple versions of control is a "local" $ITT$ effect, in the sense that it depends not only on the distribution of the outcomes but also on the distribution of the versions of control that holds in the population in which the RCT was performed. As a result, the generalizability issues discussed in the context of observational data remain present even when the data arise from a randomized experiment. We discuss how conditional causal effects from a principal stratification approach can aid in understanding how the results from an RCT conducted in one population may generalize to another, similar population that differs with respect to the underlying distribution of versions of control.

### *1.2 Notation*

In the following, we assume that we have a sample of independent observations from $n$ individuals participating in a randomized trial. The observations take the form of

triples $(Y_i, Z_i, D_i)$. The outcome $Y_i$ is assumed to be binary. The binary treatment assignment $Z_i$ is assumed to be randomized such that $P(Z_i = 1) = p_Z$ is known. The categorical covariate $D_i$ is observed post-randomization and represents information related to the version of treatment $Z_i$ actually received. The possible values that $D_i$ can take will vary based on the value of $Z_i$.

Causal inference will be conducted using the potential outcomes framework [Rubin, 1974]. Adjustment for the post-randomization covariate $D_i$ will be done using a principal stratification approach [Frangakis and Rubin, 2002a]. We assume that each individual has a pair of potential treatment versions $(D_i(0), D_i(1))$. We will let $\mathcal{D}_1 = \{0, 1\}$ designate the versions of treatment available to individuals assigned to the $Z = 1$ treatment arm. The potential treatment version $D_i(1) = 1$ indicates that an individual assigned to receive the intervention will receive the intervention, while $D_i(1) = 0$ indicates that an individual assigned to receive the intervention will not receive the intervention. Our framework thus allows for potential non-compliance to the active intervention. We will let $\mathcal{D}_0 = \{0, 1, 2\}$ designate the versions of control available to individuals assigned to the $Z = 0$ treatment arm. The potential treatment version $D_i(0) = 0$ indicates an individual assigned to the control arm will not receive any type of community-based treatment. The potential treatment version $D_i(0) = 1$ indicates that an individual assigned to the control arm will receive community-based intervention that is judged to be essentially equivalent to the active intervention under study. The potential treatment version $D_i(0) = 2$ indicates that an individual assigned to the control arm will receive community-based treatment that differs in some substantial way from the active intervention under study. As an example, suppose that the intervention of interest is a type of one-on-one therapy. In this case, $D_i(0) = 1$ might indicate that the individual in the control arm will receive one-on-one therapy from community providers, while $D_i(0) = 2$ might indicate that the individual will receive group-based therapy from community providers.

The presence of multiple versions in the control arm and non-compliance in the treatment arm means that the potential outcomes depend on both the treatment assignment and the version of treatment received. Thus, the potential outcomes for the $i^{th}$ individual must be written as $Y(z, d)$, a function of both $z \in \{0, 1\}$ and $d \in \mathcal{D}_z$. Conditional on a particular principal stratum, however, the version of treatment or control received is determined by the treatment assignment $z$. Because we will be

working within a principal stratification framework, we will suppress this notational dependence of the potential outcomes on the version received going forward and simply define $Y(z) \equiv Y(z, D(z))$.

The causal model can thus be parameterized in terms of two sets of parameters: the causal principal strata parameters, which determine the distribution of the principal strata, and the causal outcome parameters, which determine the distribution of the potential outcomes within each principal stratum. We will let $\omega_{d_0 d_1} = P(D(0) = d_0, D(1) = d_1)$ indicate the probability of belonging to the principal stratum $(D(0) = d_0, D(1) = d_1)$. We will let $p_{d_0 d_1 \cdot z} = P(Y(z) = 1 \mid D(0) = d_0, D(1) = d_1)$ indicate the probability of recovery (as defined by the outcome under consideration) under treatment assignment $Z = z$ for those in the principal stratum $(D(0) = d_0, D(1) = d_1)$.

### *1.3 Assumptions*

In this section, we outline the assumptions that we will make and their effect on the causal model parameters. First, we will make the three commonly used assumptions of non-interference, exchangeability due to randomization, and independence and identical distribution of observations. Non-interference states that the counterfactual quantities for the $i^{th}$ individual are not related to the counterfactual quantities for the other individuals. Exchangeability states that the counterfactual values are independent of the treatment assignment, which follows from the assumption that $Z$ is randomized. Finally, the independence and identical distribution assumption states that the joint distribution of the observed data for each individual $(Y_i, Z_i, D_i)$ can be modeled as coming from a common true distribution function $F$.

**Assumption 1** *a.)* Non-interference.
*Let* $\boldsymbol{z}$ *and* $\boldsymbol{d}$ *be vectors of treatment assignments and treatment versions for the* $n$ *individuals in the sample. Then for all* $i = 1, 2, \ldots, n$,

$$Y_i(\boldsymbol{z}, \boldsymbol{d}) = Y_i(z_i, d_i) \text{ and } D_i(\boldsymbol{z}) = D_i(z_i).$$

*b.)* Exchangeability/randomization.

*For all* $i = 1, 2, \ldots, n, z \in \{0, 1\}$, *and* $d \in \mathcal{D}_z$

$$Z_i \perp\!\!\!\perp D_i(z), Y_i(z, d).$$

*c.)* Independence.
*For all* $i = 1, 2, \ldots, n$,

$$(Y_i, D_i, Z_i) \overset{iid}{\sim} F$$

*for some distribution* $F$.

Later, we will examine in more detail the form that the distribution $F$ takes.

In addition to these standard assumptions, we will make the following extended consistency assumption, which establishes the connection between the counterfactual values and the observed data.

**Assumption 2** Extended Consistency assumption.
*For all* $i$,

$$D_i = D_i(0) \times (1 - Z_i) + D_i(1) \times Z_i$$

*and*

$$\begin{aligned} Y_i = {} & Y_i(0,0) \times (1 - Z_i)\mathbb{1}(D_i = 0) + Y_i(0,1) \times (1 - Z_i)\mathbb{1}(D_i = 1) + Y_i(0,2) \times (1 - Z_i)\mathbb{1}(D_i = 2) \\ & + Y_i(1,0) \times Z_i\mathbb{1}(D_i = 0) + Y_i(1,1) \times Z_i\mathbb{1}(D_i = 1) \end{aligned}$$

This extended consistency assumption differs from the typical consistency assumption in that the connection between the observed outcome and the potential outcomes depends both on the observed treatment assignment and the observed version of treatment received. Hence, if we are unable to observe the version of treatment received by an individual then we cannot determine which potential outcome we have observed, even if the treatment assignment is known. As mentioned in Section 3.2, we will define $Y_i(z) \equiv Y_i(z, D(z))$ to simplify notation.

We will make a positivity assumption for the principal strata probabilites $\omega_{01}$ and

$\omega_{21}$.

**Assumption 3** Positivity assumption.
*The principal strata probabilities $\omega_{01}$ and $\omega_{21}$ are both non-zero.*

We will also make the following assumption about the potential versions of treatment in the $Z = 0$ and $Z = 1$ arms.

**Assumption 4** Self-motivated treatment assumption.
*For all $i$,*

$$D_i(0) \geq 1 \implies D_i(1) = 1.$$

This assumption is based on the idea that the community-based treatment chosen by an individual in the TAU control group corresponds to treatment that the individual must seek out on their own. It is reasonable, then, to assume that if an individual would seek out a type of active treatment on their own, then they would similarly accept a treatment being offered to them as part of a study. We note that this assumption is similar to the commonly invoked monotonicity assumption in the binary non-compliance literature [Imbens and Rubin, 1997] [Angrist et al., 1996]. The self-motivated treatment assumption has the effect of reducing the number of principal strata from six to four. Table 1.1 summarizes the principal strata and the corresponding probability parameter for each stratum allowed under this assumption. Going forward, we will let $\mathcal{D}$ stand for the subset of pairs $(d_0, d_1)$ in the Cartesian product $\mathcal{D}_0 \times \mathcal{D}_1$ that are permissible under the self-motivated treatment assumption.

Finally, we will have reason to consider situations where the following exclusion restriction assumption holds.

**Assumption 5** Exclusion restriction assumption.
*If $d_0 = d_1$, then $p_{d_0 d_1 \cdot 1} = p_{d_0 d_1 \cdot 0} \equiv p_{d_0 d 1}$.*

This assumption states that the assignment mechanism has no effect on the potential outcomes $Y(0)$ and $Y(1)$ except through the version of treatment actually received.

This assumption is related to instrumental variable analyses often seen in the econometrics literature [Angrist et al., 1996]. When appropriate, it allows for improved estimation and inference. However, it is not necessary for all of the results that follow.

To summarize, going forward we will assume that Assumptions 1-4 hold in all of the scenarios we consider. For certain scenarios, we will also assume that Assumption 5, the *exclusion restriction* assumption, holds. When this assumption is needed, it will be explicitly mentioned.

| | $D(1) = 0$ | $D(1) = 1$ |
|---|---|---|
| $D(0) = 0$ | $\omega_{00}$ | $\omega_{01}$ |
| $D(0) = 1$ | - | $\omega_{11}$ |
| $D(0) = 2$ | - | $\omega_{21}$ |

Table 1.1: Principal strata and their associated probability parameters. Dashes indicate the principal strata that are ruled out by the self-motivated treatment assumption.

## *1.4 Effect of Control Group Heterogeneity on Generalizability of the $ITT$ Effect*

An important aspect of interpreting the conclusions from an RCT is evaluating the extent to which the effects observed in a sample from one population can be used to inform policy decisions in other populations of interest. This is referred to as generalizability or transferability in the causal inference literature [Hernán and VanderWeele, 2011]. In this section, we examine how the presence of multiple versions of control can affect the generalizability of the $ITT$ effect. This examination of the generaliziblity of the $ITT$ effect under multiple versions of control will motivate the causal estimands on which we will focus.

Continuing with the motivating example, suppose that a group of researchers at a research university in City A conducts an RCT of a behavioral intervention for children with autism using a sample drawn from the population of the metropolitan area in which the university is located. Because of the lack of a placebo-type control condition, the researchers use a TAU control condition. In addition, suppose that

City A has well-established community-based treatment options, so that control group heterogeneity of the type we have considered here is present. The authors of the study estimate a positive $ITT$ effect, which we designate as $ITT_A$ to emphasize that it comes from a study using a sample drawn from the population in City A. Now suppose that a practitioner or decision maker in City B is interested in implementing a new treatment program for children with autism in their city. Reading the results of the study done in City A, they notice that the reported demographic characteristics of the sample of children with autism from City A are nearly identical to the characteristics of children with autism in City B. Given these similarities, it is natural to ask whether the estimated effect $ITT_A$ can be used to inform policy in City B. In other words, can the effect estimated in this population (defined mainly by geographic area) be generalized to this other, highly similar population. This question is highly relevant, since many RCTs for behavioral interventions (and other classes of interventions as well) use samples drawn from a single geographic area. Of course, these studies would be of limited utility if the goal were just to understand the effect of the intervention when applied in a particular geographic area. Conducting a new study in every geographic location in which an intervention might be implemented would be costly and time consuming. Ultimately, researchers would like to be able to use the results from a limited number of studies to understand the effect of the intervention on a collection of similar populations, even if those populations happen to be geographically distinct. As discussed in Hernán and VanderWeele [2011], $ITT_A$ is a valid estimate of the $ITT$ effect *within City A*, even in the presence of control group heterogeneity. To examine how applicable $ITT_A$ is to understanding the $ITT$ effect that might be found in a study conducted in City B, we can consider the following decomposition of the $ITT$ effect in the presence of control group heterogeneity

$$
\begin{aligned}
ITT &= E[Y \mid Z = 1] - E[Y \mid Z = 0] \\
&= E[Y(1)] - E[Y(0)] \\
&= \sum_{(d_0,d_1)\in\mathcal{D}} P[Y(1) = 1, D(0) = d_0, D(1) = d_1] - P[Y(0) = 1, D(0) = d_0, D(1) = d_1] \\
&= \sum_{(d_0,d_1)\in\mathcal{D}} p_{d_0d_1\cdot 1}\omega_{d_0d_1} - p_{d_0d_1\cdot 0}\omega_{d_0d_1} \\
&= \sum_{(d_0,d_1)\in\mathcal{D}} (p_{d_0d_1\cdot 1} - p_{d_0d_1\cdot 0})\omega_{d_0d_1} \\
&\equiv \sum_{(d_0,d_1)\in\mathcal{D}} \Delta_{d_0d_1}\omega_{d0d_1}
\end{aligned}
$$

where we let $\Delta_{d_0d_1}$ represent the within-strata $ITT$ effect $p_{d_0d_1\cdot 1} - p_{d_0d_1\cdot 0}$. Thus the $ITT$ effect for a particular population can be expressed as a weighted average of the within-strata effects $\Delta_{d_0d_1}$, where the weights are given by the distribution $\boldsymbol{\omega} = (\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21})$ of principal strata for that population. This decomposition makes clear how useful estimating $ITT_A$ will be as a way of making decisions about $ITT_B$. Even if the populations of children with autism in City A and City B are so similar that each of the within-strata causal effects $\Delta_{d_0d_1}$ are equal between the two cities, $ITT_A$ will likely not generalize to $ITT_B$ unless the distribution of principal strata (i.e. the pattern of control group heterogeneity) is the same between the two cities. By looking at stratum-specific causal effects, we can gain insight into how the results from a study performed in one population might generalize to populations that differ in the availability of community-based treatment options but are otherwise similar. The causal estimands we will focus on are the stratum-specific effects $\Delta_{01}$ and $\Delta_{21}$:

$$
\begin{aligned}
\Delta_{01} &= P(Y(1) = 1 \mid D(0) = 0, D(1) = 1) - P(Y(0) = 1 \mid D(0) = 0, D(1) = 1) \\
&= p_{01\cdot 1} - p_{01\cdot 0},
\end{aligned}
$$

$$\begin{aligned}\Delta_{21} &= P(Y(1) = 1 \mid D(0) = 2, D(1) = 1) - P(Y(0) = 1 \mid D(0) = 2, D(1) = 1) \\ &= p_{21\cdot 1} - p_{21\cdot 0}.\end{aligned}$$

Similar to the Complier Average Causal Effect (CACE) in the binary non-compliance literature, the causal estimand $\Delta_{01}$ corresponds to the causal effect among those who would receive the active intervention if assigned to it and no intervention otherwise. This causal estimand is useful for understanding the effect of the intervention if implemented in an area without well-established community-based treatment options. The causal estimand $\Delta_{21}$ corresponds to the causal effect among those who would receive the active intervention if assigned to it and some alternative form of treatment if assigned to the control group. This causal estimand is useful for understanding the effect of the intervention if implemented in an area with well-established community-based treatment options that are distinct from the intervention being studied.

### *1.5 Identifiability of $\Delta_{01}$ and $\Delta_{21}$*

Under the assumptions outlined in Section 1.3, the causal estimands $\Delta_{01}$ and $\Delta_{21}$ lack point identifiability. Lack of point identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$ follows from a disparity between the number of causal parameters in the model for the counterfactual values and the number of statistical parameters in the model for the observed data. In this section, we demonstrate the identifiability issue with a numerical example. We then examine which of the causal model parameters are identified and which are not, both under Assumptions 1-4 and Assumptions 1-5.

We will parameterize the observed-data distribution in terms of the joint probabilities

$$q_{dy\cdot z} = P(Y = y, D = d \mid Z = z).$$

which fully determine the distribution of the observed data. In all that follows, we will parameterize the observed-data distribution in terms of the eight element vector $\boldsymbol{q} = (q_{01\cdot 0}, q_{11\cdot 0}, q_{10\cdot 0}, q_{20\cdot 0}, q_{21\cdot 0}, q_{00\cdot 1}, q_{01\cdot 1}, q_{11\cdot 1})^T$. This eight element vector

completely parameterizes the observed-data distribution since the missing elements $q_{00\cdot 0}$ and $q_{10\cdot 1}$ can be written in terms of elements of $\boldsymbol{q}$. The elements $q_{dy\cdot z}$ of $\boldsymbol{q}$ can be easily estimated by their sample counterparts

$$\hat{q}_{dy\cdot z} = \frac{\sum_{i=1}^{n} \mathbb{1}(D_i = d)\mathbb{1}(Y_i = y)\mathbb{1}(Z_i = z)}{\sum_{i=1}^{n} \mathbb{1}(Z_i = z)}$$

where $\mathbb{1}(\cdot)$ represents the indicator function that is equal to 1 if the condition in the parentheses is satisfied and 0 otherwise. We will let $\hat{\boldsymbol{q}} = (\hat{q}_{01\cdot 0}, \hat{q}_{11\cdot 0}, \hat{q}_{10\cdot 0}, \hat{q}_{20\cdot 0}, \hat{q}_{21\cdot 0}, \hat{q}_{00\cdot 1}, \hat{q}_{01\cdot 1}, \hat{q}_{11\cdot 1})^T$ represent the vector of estimated observe-data quantities.

Because the estimator $\hat{\boldsymbol{q}}$ of the observed-data quantities $\boldsymbol{q}$ will play an important role in estimation and inference for the bounds on $\Delta_{01}$ and $\Delta_{21}$, we describe its asymptotic behavior in the following lemma.

**Lemma 1** *The estimator $\hat{\boldsymbol{q}}$ is a consistent estimator for $\boldsymbol{q}$,*

$$\hat{\boldsymbol{q}} \xrightarrow{p} \boldsymbol{q}$$

*with asymptotic distribution given by*

$$\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q}) \xrightarrow{d} N_8(\mathbf{0}, \Sigma_q)$$

*where*

$$\Sigma_q = \begin{pmatrix} \frac{1}{p_0}\Sigma_0 & \mathbf{0} \\ \mathbf{0} & \frac{1}{p_1}\Sigma_1 \end{pmatrix}$$

*with*

$$\Sigma_0 = \begin{array}{c} \\ q_{01\cdot 0} \\ q_{11\cdot 0} \\ q_{10\cdot 0} \\ q_{20\cdot 0} \\ q_{21\cdot 0} \end{array} \begin{array}{c} \begin{array}{ccccc} q_{01\cdot 0} & q_{11\cdot 0} & q_{10\cdot 0} & q_{20\cdot 0} & q_{21\cdot 0} \end{array} \\ \begin{pmatrix} q_{01\cdot 0}(1-q_{01\cdot 0}) & -q_{01\cdot 0}q_{11\cdot 0} & -q_{01\cdot 0}q_{10\cdot 0} & -q_{01\cdot 0}q_{20\cdot 0} & -q_{01\cdot 0}q_{21\cdot 0} \\ -q_{01\cdot 0}q_{11\cdot 0} & q_{11\cdot 0}(1-q_{11\cdot 0}) & -q_{10\cdot 0}q_{11\cdot 0} & -q_{11\cdot 0}q_{20\cdot 0} & -q_{11\cdot 0}q_{21\cdot 0} \\ -q_{01\cdot 0}q_{10\cdot 0} & -q_{10\cdot 0}q_{11\cdot 0} & q_{10\cdot 0}(1-q_{10\cdot 0}) & -q_{10\cdot 0}q_{20\cdot 0} & -q_{10\cdot 0}q_{21\cdot 0} \\ -q_{01\cdot 0}q_{20\cdot 0} & -q_{11\cdot 0}q_{20\cdot 0} & -q_{10\cdot 0}q_{20\cdot 0} & q_{20\cdot 0}(1-q_{20\cdot 0}) & -q_{20\cdot 0}q_{21\cdot 0} \\ -q_{01\cdot 0}q_{21\cdot 0} & -q_{11\cdot 0}q_{21\cdot 0} & -q_{10\cdot 0}q_{21\cdot 0} & -q_{20\cdot 0}q_{21_c dot 0} & q_{21\cdot 0}(1-q_{21\cdot 0}) \end{pmatrix} \end{array}$$

*and*

$$\Sigma_1 = \begin{array}{c} \\ q_{00\cdot1} \\ q_{01\cdot1} \\ q_{11\cdot1} \end{array} \begin{array}{c} \begin{array}{ccc} q_{00\cdot1} & q_{01\cdot1} & q_{11\cdot1} \end{array} \\ \begin{pmatrix} q_{00\cdot1}(1-q_{00\cdot1}) & -q_{00\cdot1}q_{01\cdot1} & -q_{00\cdot1}q_{01\cdot1} \\ -q_{00\cdot1}q_{01\cdot1} & q_{01\cdot1}(1-q_{01\cdot1}) & -q_{01\cdot1}q_{11\cdot1} \\ -q_{00\cdot1}q_{11\cdot1} & -q_{01\cdot1}q_{11\cdot1} & q_{11\cdot1}(1-q_{11\cdot1}) \end{pmatrix} \end{array}.$$

Importantly, the consistency and asymptotic normality of the estimators $\hat{q}_{dy\cdot z}$ do not depend on the exclusion restriction assumption. Thus, the results in Lemma 1 can be used in studying the large sample behavior of the bounds estimators for $\Delta_{01}$ and $\Delta_{21}$ under either Assumptions 1-4 or Assumptions 1-5.

We now provide a numerical example demonstrating that $\Delta_{01}$ and $\Delta_{21}$ are not point identified. Our numerical example will incorporate the exclusion restriction assumption. Non-identifiability can likewise be established in cases where the exclusion restriction does not hold. In fact, the identifiability issue is arguably even more severe when the exclusion restriction assumption does not hold, since it implies the presence of additional causal parameters but no additional observed-data quantities. Letting $\boldsymbol{\theta} = (p_{00\cdot1}, p_{01\cdot1}, p_{11\cdot1}, p_{21\cdot1}, p_{00\cdot0}, p_{01\cdot0}, p_{11\cdot0}, p_{21\cdot0}, \omega_{00}, \omega_{01}, \omega_{11}, \omega_{21})^T$ represent the full vector of causal parameters, we can consider the two possible parameter vectors for the underlying causal model

$$\boldsymbol{\theta}_1 = (.03, .3, .2, .2, .03, .03, .2, .15, .1, .4, .1, .4)^T$$
$$\boldsymbol{\theta}_2 = (.03, .2, .2, .3, .03, .03, .2, .15, .1, .4, .1, .4)^T$$

which differ with respect to the causal outcome parameters $p_{01\cdot1}$ and $p_{21\cdot1}$. These causal parameter vectors can be used to explicitly compute the corresponding observed-data distributions $P_{\theta_1}$ and $P_{\theta_2}$ (see Table 1.2). Under this setup, the observed-data distributions implied by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are identical. However, $\Delta_{01}^{\theta_1} = .3 - .03 = .27$ while $\Delta_{01}^{\theta_2} = .2 - .03 = .17$, and $\Delta_{21}^{\theta_1} = .2 - .15 = .05$ while $\Delta_{21}^{\theta_2} = .3 - .15 = .15$. Since different underlying values for $\Delta_{01}$ and $\Delta_{21}$ can lead to the same observed-data distribution, we can conclude that $\Delta_{01}$ and $\Delta_{21}$ are not point identified.

Having demonstrated the identifiability issues surrounding $\Delta_{01}$ and $\Delta_{21}$, we now derive point identified upper and lower bounds on these causal estimands. To establish

| Causal Parameter | $q_{01\cdot1}$ | $q_{00\cdot1}$ | $q_{11\cdot1}$ | $q_{01\cdot0}$ | $q_{00\cdot0}$ | $q_{11\cdot0}$ | $q_{10\cdot0}$ | $q_{21\cdot0}$ |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}_1$ | 0.003 | 0.097 | 0.220 | 0.015 | 0.485 | 0.020 | 0.080 | 0.060 |
| $\boldsymbol{\theta}_2$ | 0.003 | 0.097 | 0.220 | 0.015 | 0.485 | 0.020 | 0.080 | 0.060 |

Table 1.2: Observed-data distributions implied by causal parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The agreement between the two observed-data distributions despite different underlying values for $\Delta_{01}$ and $\Delta_{21}$ demonstrates the identifiability issue.

these bounds, it will be useful to know which of the causal model parameters are point identified and which are not. This will allow us to know which parameters can be treated as fixed, and which parameters we must maximize or minimize with respect to when deriving the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$. In Appendix A, we give results for the identifiability of all causal parameters described in Section 3.2, both under Assumptions 1-4 and under Assumptions 1-5. We summarize the identifiability results for the causal model parameters relevant to deriving the bounds on $\Delta_{01}$ and $\Delta_{21}$ in the following lemma.

**Lemma 2** *a.) Under both sets of assumptions given in Section 1.3, the strata membership probabilities* $\boldsymbol{\omega} = (\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21})$ *are point identified and expressible in terms of the observed data quantities* $q_{dy\cdot z}$ *as follows. For* $d_0 = 1, 2$ *and* $d_1 = 1$*,*

$$\omega_{d_0 1} = q_{d0\cdot0} + q_{d1\cdot0}.$$

*For* $d_0 = 0$ *and* $d_1 = 0$*,*

$$\omega_{00} = q_{00\cdot1} + q_{01\cdot1}.$$

*For* $d_0 = 0$ *and* $d_1 = 1$*,*

$$\omega_{01} = 1 - (q_{10\cdot0} + q_{11\cdot0} + q_{20\cdot0} + q_{21\cdot0} + q_{01\cdot1} + q_{00\cdot1}).$$

*b.) Under Assumptions 1-4, the causal outcome parameter* $p_{21\cdot0}$ *is point identified and expressible in terms of observed data quantities as follows*

$$p_{21\cdot0} = \frac{q_{21\cdot0}}{\omega_{21}}.$$

*The causal outcome parameters $p_{00\cdot 0}, p_{01\cdot 0}, p_{01\cdot 1}, p_{11\cdot 1}$, and $p_{21\cdot 1}$ are unidentified under Assumptions 1-4.*
*c.) Under Assumptions 1-5, the causal outcome parameters $p_{01\cdot 0}$ and $p_{21\cdot 0}$ are point identified and expressible in terms of observed data quantities as follows*

$$p_{21\cdot 0} = \frac{q_{21\cdot 0}}{\omega_{21}}$$
$$p_{01\cdot 0} = \frac{q_{01\cdot 0} - q_{01\cdot 1}}{\omega_{01}}.$$

*The causal outcome parameters $p_{01\cdot 1}$ and $p_{21\cdot 1}$ are unidentified under Assumptions 1-5.*

Estimators for the point identified parameters can be constructed by replacing the observed data quantities $q_{dy\cdot z}$ on the right hand side of the identifying equations with their sample-based equivalents $\hat{q}_{dy\cdot z}$. These estimators will appear in the bounds estimators for $\Delta_{01}$ and $\Delta_{21}$. Their asymptotic behavior is described in the following lemma.

**Lemma 3** *a.) Under both sets of assumptions given in Section 1.3, the estimator $\hat{\boldsymbol{\omega}}$ is a consistent and asymptotically normal estimator of $\boldsymbol{\omega}$.*
*b.) Under Assumptions 1-4, the estimator $\hat{p}_{21\cdot 0}$ is a consistent and asymptotically normal estimator of $p_{21\cdot 0}$.*
*c.) Under Assumptions 1-5, the estimators $\hat{p}^{ER}_{01\cdot 0}$ and $\hat{p}^{ER}_{21\cdot 0}$ are consistent and asymptotically normal estimators for $p_{01\cdot 0}$ and $p_{21\cdot 0}$, respectively, where the superscript $ER$ indicates that these estimators are based on the expressions for $p_{01\cdot 0}$ and $p_{21\cdot 0}$ derived under the exclusion restriction assumption.*

The proofs of Lemmas 2 and 3 appear in Appendix A. We note that the identifiability results for the principal strata membership probabilities follow chiefly from the consistency and self-motivated treatment assumptions. As a consequence of this, the identifiability results for the principal strata probabilities hold regardless of whether the Exclusion Restriction assumption is made.

### *1.6 Linear Programming Bounds for $\Delta_{01}$ and $\Delta_{21}$*

Although the causal parameters $\Delta_{01}$ and $\Delta_{21}$ are not point identified, we show in this section how they can be bounded based on observed-data quantities. The simplest

approach would be to simply replace any unidentified causal parameter in $\Delta_{01}$ and $\Delta_{21}$ with its corresponding natural bound. Under Assumptions 1-5, for instance, this would give an upper bound on $\Delta_{01}$ of $1 - p_{01\cdot0}$ and a lower bound of $-p_{01\cdot0}$. Since $p_{01\cdot0}$ is point identified under Assumptions 1-5, the upper and lower bounds constructed in this way are estimable from the data. While these bounds would likely be an improvement on the natural bounds of $[-1, 1]$, they are not desirable for several reasons. Similar to the bounds derived in Manski [1990], the bounds derived in this way would necessarily have to cover 0, and thus the direction of the effect could never be determined based on them. In addition, these bounds fail to exploit all of the information available in the data about the identified regions of $\Delta_{01}$ and $\Delta_{21}$. An alternative approach that does take advantage of additional information in the data is to determine the upper and lower bounds as the solutions to a linear programming problem. In a linear programming problem, the goal is to maximize (or minimize) a linear function of one or more unknown variables, known as the objective function, subject to a set of constraints on those variables. The constraints can be in the form of inequalities or equalities, but must again be linear in the unknown variables. Linear programming techniques are a common approach to finding bounds on partially identified causal parameters based on the available data [Balke and Pearl, 1997] [Cheng and Small, 2006] [Yang and Small, 2016].

### *1.6.1 Linear Programming Bounds without Exclusion Restriction*

We wish to find minimum and maximum values for $\Delta_{01} = p_{01\cdot1} - p_{01\cdot0}$ and $\Delta_{21} = p_{21\cdot1} - p_{21\cdot0}$ under Assumptions 1-4. From Lemma 2, we know that the unidentified causal parameters under Assumptions 1-4 are $p_{01\cdot0}, p_{01\cdot1}$ and $p_{21\cdot1}$. We also have the following relationships between the observed data quantities $\boldsymbol{q}$ and these unidentified causal outcome parameters

$$q_{11\cdot1} = \omega_{01}p_{01\cdot1} + \omega_{11}p_{11\cdot1} + \omega_{21}p_{21\cdot1}$$
$$q_{01\cdot0} = \omega_{00}p_{00\cdot0} + \omega_{01}p_{01\cdot0}.$$

Thus, the minimization/maximization of $\Delta_{01}$ and $\Delta_{21}$ subject to the constraints implied by Assumptions 1-4 can be viewed as two linear programming problems with

objective functions $p_{01\cdot 1} - p_{01\cdot 0}$ and $p_{21\cdot 1}$. We now outline how this linear programming problem can be solved to obtain the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$. Because the quantities $\boldsymbol{q}, \boldsymbol{\omega}$, and $p_{21\cdot 0}$ are point identified under Assumptions 1-4, they can be treated as known constants. By considering the extreme values that the unidentified causal parameters $p_{01\cdot 1}, p_{01\cdot 0}, p_{00\cdot 1}$ and $p_{21\cdot 1}$ can take, we obtain the following unrestrained bounds on $p_{01\cdot 0}$ and $p_{21\cdot 0}$

$$
\begin{aligned}
p^U_{01\cdot 1} &= \frac{q_{11\cdot 1}}{\omega_{01}} \\
p^L_{01\cdot 1} &= \frac{q_{11\cdot 1} - \omega_{11} - \omega_{21}}{\omega_{01}} \\
p^U_{01\cdot 0} &= \frac{q_{01\cdot 0}}{\omega_{01}} \\
p^L_{01\cdot 0} &= \frac{q_{01\cdot 0} - \omega_{00}}{\omega_{01}} \\
p^U_{21\cdot 1} &= \frac{q_{11\cdot 1}}{\omega_{21}} \\
p^L_{21\cdot 1} &= \frac{q_{11\cdot 1} - \omega_{01} - \omega_{11}}{\omega_{21}}.
\end{aligned}
$$

While these are valid upper and lower bounds on their respective causal parameters, they are unrestrained in the sense that they may be greater than 1 (in the case of the upper bounds) or less than 0 (in the case of the lower bounds). The linear programming-based upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ can be expressed in terms of these unrestrained bounds

$$
\begin{aligned}
\Delta^U_{01} &= \min\{1, p^U_{01\cdot 1}\} - \max\{0, p^L_{01\cdot 0}\} \\
\Delta^L_{01} &= \max\{0, p^L_{01\cdot 1}\} - \min\{1, p^U_{01\cdot 0}\} \\
\Delta^U_{21} &= \min\{1, p^U_{21\cdot 1}\} - p_{21\cdot 0} \\
\Delta^L_{21} &= \max\{0, p^L_{21\cdot 1}\} - p_{21\cdot 0}
\end{aligned}
$$

with corresponding estimators

$$\begin{aligned}
\hat{\Delta}_{01}^{U} &= \min\{1, \hat{p}_{01\cdot 1}^{U}\} - \max\{0, \hat{p}_{01\cdot 0}^{L}\}, \\
\hat{\Delta}_{01}^{L} &= \max\{0, \hat{p}_{01\cdot 1}^{L}\} - \min\{1, \hat{p}_{01\cdot 0}^{U}\}, \\
\hat{\Delta}_{21}^{U} &= \min\{1, \hat{p}_{21\cdot 1}^{U}\} - \hat{p}_{21\cdot 0}, \\
\hat{\Delta}_{21}^{L} &= \max\{0, \hat{p}_{21\cdot 1}^{L}\} - \hat{p}_{21\cdot 0}.
\end{aligned}$$

### *1.6.2 Linear Programming Bounds with Exclusion Restriction*

We now show how to find minimum and maximum values for $\Delta_{01} = p_{01\cdot 1} - p_{01\cdot 0}$ and $\Delta_{21} = p_{21\cdot 1} - p_{21\cdot 0}$ under Assumptions 1-5. From Lemma 2, the unidentified causal parameters under Assumptions 1-5 are $p_{01\cdot 1}$ and $p_{21\cdot 1}$. We additionally have the following relationship between these unidentified causal outcome parameters and the observed data

$$q_{11\cdot 1} - q_{11\cdot 0} = \omega_{01} p_{01\cdot 1} + \omega_{21} p_{21\cdot 1}.$$

Thus, the minimization/maximization of $\Delta_{01}$ and $\Delta_{21}$ under the constraints implied by Assumptions 1-5 can be viewed as two linear programming problems with objective functions $p_{01\cdot 1}$ and $p_{21\cdot 1}$. We now outline how this linear programming problem can be solved to obtain upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$. Because $\boldsymbol{q}$ and $\boldsymbol{\omega}$ are point identified under Assumptions 1-5, they can be treated as known constants in the linear programming problem. By considering the extreme values that $p_{01\cdot 1}$ and $p_{21\cdot 1}$ can assume, we have the following unrestrained lower and upper bounds on $p_{01\cdot 1}$ and $p_{21\cdot 1}$ under Assumptions 1-5.

$$
\begin{aligned}
p_{01\cdot1}^{L,ER} &= \frac{q_{11\cdot1} - q_{11\cdot0} - \omega_{21}}{\omega_{01}} \\
p_{01\cdot1}^{U,ER} &= \frac{q_{11\cdot1} - q_{11\cdot0}}{\omega_{01}} \\
p_{21\cdot1}^{L,ER} &= \frac{q_{11\cdot1} - q_{11\cdot0} - \omega_{01}}{\omega_{21}} \\
p_{21\cdot1}^{U,ER} &= \frac{q_{11\cdot1} - q_{11\cdot0}}{\omega_{21}}
\end{aligned}
$$

where the superscript $ER$ emphasizes that these bounds are dependent on the exclusion restriction assumption. While these upper and lower bounds are valid bounds for their respective parameters, they are unrestrained in the sense that they may be greater than 1 (in the case of upper bounds) or less than 0 (in the case of lower bounds). Thus, bounds on $\Delta_{01}$ and $\Delta_{21}$ based on these unrestrained bounds may not be the tightest bounds possible. The linear programming-based bounds on $\Delta_{01}$ and $\Delta_{21}$ give the tightest bounds and are given by

$$
\begin{aligned}
\Delta_{01}^{U,ER} &= \min\{1, p_{01\cdot1}^{U,ER}\} - p_{01\cdot0} \\
\Delta_{01}^{L,ER} &= \max\{0, p_{01\cdot1}^{L,ER}\} - p_{01\cdot0} \\
\Delta_{21}^{U,ER} &= \min\{1, p_{21\cdot1}^{U,ER}\} - p_{21\cdot0} \\
\Delta_{21}^{L,ER} &= \max\{0, p_{21\cdot1}^{L,ER}\} - p_{21\cdot0}.
\end{aligned}
$$

with corresponding estimators

$$
\begin{aligned}
\hat{\Delta}_{01}^{U,ER} &= \min\{1, \hat{p}_{01\cdot1}^{U,ER}\} - \hat{p}_{01\cdot0} \\
\hat{\Delta}_{01}^{L,ER} &= \max\{0, \hat{p}_{01\cdot1}^{L,ER}\} - \hat{p}_{01\cdot0} \\
\hat{\Delta}_{21}^{U,ER} &= \min\{1, \hat{p}_{21\cdot1}^{U,ER}\} - \hat{p}_{21\cdot0} \\
\hat{\Delta}_{21}^{L,ER} &= \max\{0, \hat{p}_{21\cdot1}^{L,ER}\} - \hat{p}_{21\cdot0}
\end{aligned}
$$

where again the superscript $ER$ is used to emphasize that these upper and lower bounds are based on Assumptions 1-5. It can be shown that the bounds derived under Assumptions 1-5 are tighter than the bounds derived under Assumptions 1-4,

so that being able to make the exclusion restriction does indeed provide an advantage. In a later section, we discuss model checks that can be performed to see whether the observed data violate the exclusion restriction assumption.

The types of bounds on $p_{01\cdot 1}$ and $p_{21\cdot 1}$ that can result from the set of constraints can be visualized in two dimensions as in Figure 1.2. The different combinations shown in Figure 1.2 demonstrate how narrower bounds for one of either $p_{01\cdot 1}$ or $p_{21\cdot 1}$ necessarily implies wider bounds for the other parameter. In particular, if the identified region for one of these parameters is contained within the open interval $(0, 1)$, then the identified region for the other parameter must be the full interval $[0, 1]$.

### *1.6.3 Relative Degrees of Identifiability of $\Delta_{01}$ and $\Delta_{21}$*

We now introduce some terms to better clarify the differing levels of information available in the data in the partial identifiability setting we are considering. In this case, there are natural bounds on the quantities of interest. For the probabilities $p_{d_0 d_1 \cdot z}$ these natural bounds are 0 and 1, while for the causal estimands $\Delta_{01}$ and $\Delta_{21}$ the natural bounds are -1 and 1. We will refer to a parameter as fully unidentified if the identified subset corresponds to the natural bounds. We will refer to a parameter as asymmetrically partially identified if the identified subset contains one of the natural bounds. For instance, a probability parameter with identified subset equal to $[0, .4)$ would be asymmetrically partially identified, since the lower bound corresponds to the natural lower bound of 0. We will refer to a parameter as partially identified if its identified subset contains neither of the natural bounds, e.g. the interval $[0.4, 0.6]$ for a probability parameter, or $[-.4, .2]$ for a difference in probabilities. Finally, for parameters that can take on both positive and negative values, we will use the term informative to refer to an identified subset that does not include 0 (and thus can be used to determine the direction of an effect).

Under both sets of assumptions we have considered, the causal estimands $\Delta_{01}$ and $\Delta_{21}$ have the potential to be either fully unidentified, asymmetrically partially identified, or partially identified. However, certain degrees of identifiability are arguably more or less likely to occur in practice depending on the set of assumptions made in the analysis. Under Assumptions 1-4, the causal estimand $\Delta_{01}$ arguably has the greatest potential to be fully unidentified, due to the fact that both causal

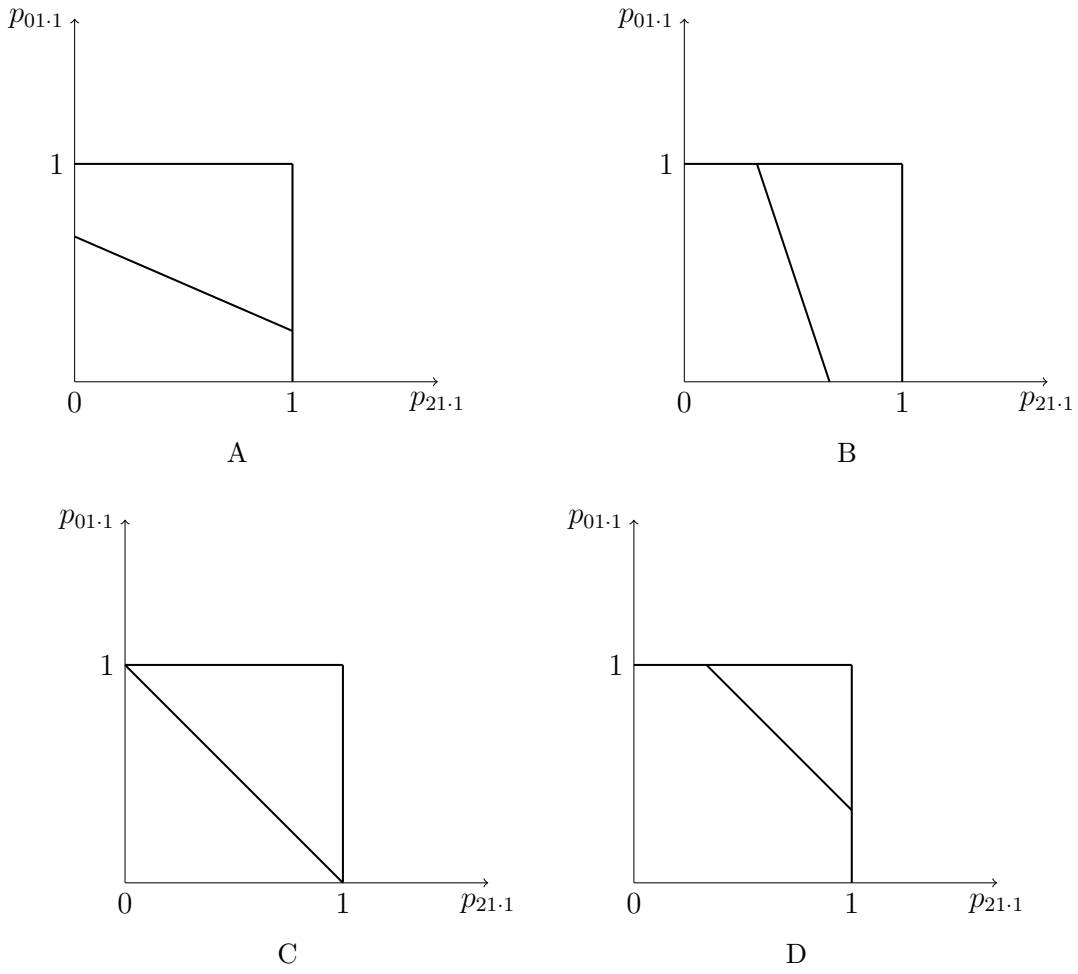



Figure 1.2: Visualization of the linear constraints on $p_{01\cdot1}$ and $p_{21\cdot1}$ under Assumptions 1-5. The panels show the examples of possible combinations of identifiability. In Panel A, $p_{01\cdot1}$ is partially identified while $p_{21\cdot1}$ is fully unidentified. In Panel B, $p_{21\cdot1}$ is partially identified while $p_{01\cdot1}$ is fully unidentified. In Panel C, both $p_{01\cdot1}$ and $p_{21\cdot1}$ are fully unidentified. In Panel D, both $p_{01\cdot1}$ and $p_{21\cdot1}$ are asymmetrically partially identified.

parameters $p_{01\cdot1}$ and $p_{01\cdot0}$ are unidentified under these assumptions. In contrast, $\Delta_{21}$ is arguably unlikely to be fully unidentified under Assumptions 1-4 since $p_{21\cdot0}$ is point identified. Similarly, under Assumptions 1-5 both $\Delta_{01}$ and $\Delta_{21}$ are unlikely to be fully unidentified in practice since both $p_{01\cdot0}$ and $p_{21\cdot0}$ are point identified under this set of assumptions.

The scenario of greatest interest for researchers is the one in which $\Delta_{01}$ and $\Delta_{21}$ are not just partially identified, but the bounds on them are actually informative as to the direction of the effect as well. Informative bounds seem least likely to be obtained in practice under Assumptions 1-4, due to the wider bounds that result from lack of the exclusion restriction assumption. In particular, we would expect informative bounds for $\Delta_{01}$ under Assumptions 1-4 to be unlikely due to the fact that both of the causal outcome parameters $p_{01\cdot1}$ and $p_{01\cdot0}$ are not point unidentified under these assumptions. Under Assumptions 1-5, informative bounds can be obtained on both $\Delta_{01}$ and $\Delta_{21}$ in situations such as the one shown in Panel D of Figure 1.2. On the other hand, the fact that the width of the bounds for a given stratum-specific estimand is inversely proportional to the probability of belonging to that stratum suggests that informative bounds are only likely to be obtained if a relatively large percentage of the population belongs to that stratum. As a result, we would not expect to be able to obtain relatively narrow, informative bounds on $\Delta_{01}$ and $\Delta_{21}$ simultaneously.

### *1.6.4 Consistency of the Bounds Estimators*

We now show that the estimators for the bounds on $\Delta_{01}$ and $\Delta_{21}$ derived under Assumptions 1-4 and Assumptions 1-5 are consistent for the true bounds. To begin, we note that the upper and lower bounds derived under Assumptions 1-4 can be written as

$$
\begin{aligned}
\Delta_{01}^{U} &= p_{01\cdot1}^{U} \times \mathbb{1}(p_{01\cdot1}^{U} < 1) + \mathbb{1}(p_{01\cdot1}^{U} \geq 1) - p_{01\cdot0}^{L} \times \mathbb{1}(p_{01\cdot0}^{L} > 0) \\
\Delta_{01}^{L} &= p_{01\cdot1}^{L} \times \mathbb{1}(p_{01\cdot1}^{L} > 0) - p_{01\cdot0}^{U} \times \mathbb{1}(p_{01\cdot0}^{U} < 1) - \mathbb{1}(p_{01\cdot0}^{U} \geq 1) \\
\Delta_{21}^{U} &= p_{21\cdot1}^{U} \times \mathbb{1}(p_{21\cdot1}^{U} < 1) + \mathbb{1}(p_{21\cdot1}^{U} \geq 1) - p_{21\cdot0} \\
\Delta_{21}^{L} &= p_{21\cdot1}^{L} \times \mathbb{1}(p_{21\cdot1}^{L} > 0) - p_{21\cdot0}
\end{aligned}
$$

and the upper and lower bounds under Assumptions 1-5 can be written as

$$\Delta_{d1}^{U,ER} = p_{d1\cdot1}^{U,ER} \times \mathbb{1}(p_{d1\cdot1}^{U,ER} < 1) + \mathbb{1}(p_{d1\cdot1}^{U,ER} \geq 1) - p_{d1\cdot0},$$
$$\Delta_{d1}^{L,ER} = p_{d1\cdot1}^{L,ER} \times \mathbb{1}(p_{d1\cdot1}^{L,ER} > 0) - p_{d1\cdot0}$$

for $d \in \{0, 2\}$.

The asymptotic behavior of the linear programming-based bounds estimators thus depends on the asymptotic behavior of the estimators of the unrestrained bounds on $p_{01\cdot0}, p_{01\cdot1}$ and $p_{21\cdot1}$ mentioned earlier. We will make use of the following lemmas describing the asymptotic behavior of these estimators.

**Lemma 4** *Under Assumptions 1-4, the estimators* $\hat{p}_{01\cdot1}^{U}, \hat{p}_{01\cdot1}^{L}, \hat{p}_{01\cdot0}^{U}, \hat{p}_{01\cdot0}^{L}, \hat{p}_{21\cdot1}^{U}, \hat{p}_{21\cdot1}^{L}$ *are consistent and asymptotically normal estimators of the upper and lower bounds* $p_{01\cdot1}^{U}, p_{01\cdot1}^{L}$, $p_{01\cdot0}^{U}, p_{01\cdot0}^{L}, p_{21\cdot1}^{U}, p_{21\cdot1}^{L}$, *respectively.*

**Lemma 5** *Under Assumptions 1-5, the estimators* $\hat{p}_{01\cdot1}^{U}, \hat{p}_{01\cdot1}^{L}, \hat{p}_{21\cdot1}^{U}, \hat{p}_{21\cdot1}^{L}$ *are consistent and asymptotically normal estimators of the upper and lower bounds* $p_{01\cdot1}^{U}, p_{01\cdot1}^{L}, p_{21\cdot1}^{U}, p_{21\cdot1}^{L}$, *respectively.*

The proofs for these lemmas are given in Appendix A. We now state the consistency results for the bounds estimators.

**Theorem 1** *a.) Under Assumptions 1-4,* $\hat{\Delta}_{d1}^{t}$ *is a consistent estimator for* $\Delta_{d1}^{t}$

$$\hat{\Delta}_{d1}^{t} \xrightarrow{p} \Delta_{d1}^{t}$$

*for* $d \in \{0, 2\}$ *and* $t \in \{L, U\}$.

*b.) Under Assumptions 1-5,* $\hat{\Delta}_{d1}^{t,ER}$ *is a consistent estimator for* $\Delta_{d1}^{t,ER}$

$$\hat{\Delta}_{d1}^{t,ER} \xrightarrow{p} \Delta_{d1}^{t,ER}$$

*for* $d \in \{0, 2\}$ *and* $t \in \{L, U\}$.

*Proof.* a.) Under Assumptions 1-4, the estimators of the upper and lower bounds on

$\Delta_{01}$ can be written as

$$\hat{\Delta}_{01}^{L} = \hat{p}_{01\cdot 1}^{L} \times \mathbb{1}(\hat{p}_{01\cdot 1}^{L} > 0) - \hat{p}_{01\cdot 0}^{U} \times \mathbb{1}(\hat{p}_{01\cdot 0}^{U} < 1) + \mathbb{1}(\hat{p}_{01\cdot 0}^{U} \geq 1),$$
$$\hat{\Delta}_{01}^{U} = \hat{p}_{01\cdot 1}^{U} \times \mathbb{1}(\hat{p}_{01\cdot 1}^{U} < 1) + \mathbb{1}(\hat{p}_{01\cdot 1} \geq 1) - \hat{p}_{01\cdot 0}^{L} \times \mathbb{1}(\hat{p}_{01\cdot 0}^{L} > 0).$$

Similarly, the estimators of the upper and lower bounds on $\Delta_{21}$ can be written as

$$\hat{\Delta}_{21}^{U} = \hat{p}_{21\cdot 1}^{U} \times \mathbb{1}(\hat{p}_{21\cdot 1}^{U} < 1) + \mathbb{1}(\hat{p}_{21\cdot 1} \geq 1) - \hat{p}_{21\cdot 0},$$
$$\hat{\Delta}_{21}^{L} = \hat{p}_{21\cdot 1}^{L} \times \mathbb{1}(\hat{p}_{21\cdot 1}^{L} > 0) - \hat{p}_{21\cdot 0}.$$

By Lemma 4, each quantity on the right hand sides of these equations individually converges in probability to its population parameter equivalent. Thus by Slutsky's theorem we can conclude that the full expression on the right hand side converges in probability to the equivalent expression with the estimators replaced by their population parameter equivalents, namely the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ derived under Assumptions 1-4.

b.) Under Assumptions 1-5, the estimators of the upper and lower bounds on $\Delta_{01}$ and $\Delta_{02}$ can be written as

$$\hat{\Delta}_{d1}^{U,ER} = \hat{p}_{d1\cdot 1}^{U,ER} \times \mathbb{1}(\hat{p}_{d1\cdot 1}^{U,ER} < 1) + \mathbb{1}(\hat{p}_{d1\cdot 1}^{U,ER} \geq 1) - \hat{p}_{d1\cdot 0},$$
$$\hat{\Delta}_{d1}^{L,ER} = \hat{p}_{d1\cdot 0}^{L,ER} \times \mathbb{1}(\hat{p}_{d1\cdot 0}^{L,ER} > 0) - \hat{p}_{d1\cdot 0}$$

for $d \in \{0, 2\}$. By Lemma 5, each quantity on the right hand sides of these equations converges in probability to its respective population parameter. Thus, by Slutsky's theorem the entire expression on the right hand side converges in probability to its population equivalent, namely the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ derived under Assumptions 1-5. ■

### *1.6.5 Inference for Linear Programming-based Bounds*

Inference for the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ can be performed in several ways. Bootstrap-based approaches to inference are a common approach in the partial identifiability literature [Cheng and Small, 2006, Horowitz and Manski, 2000, Imbens and Manski, 2004] because the large sample distributions of estimators for bounds

are often cumbersome to deal with. We will consider the use of the non-parameteric bootstrap [Efron, 1979] as one method for performing inference. A $100 \times (1-\alpha)\%$ Bonferroni-type confidence interval for the identified region $[\Delta_{d1}^L, \Delta_{d1}^U], d \in \{0, 2\}$ can be constructed by obtaining a bootstrapped sample of bounds estimators, constructing separate $100 \times (1-\alpha/2)\%$ confidence intervals for the upper and lower bounds based on the quantiles of the bootstrap distributions, and then using the lower end of the confidence interval for the lower bound and the upper end of the confidence interval for the upper bound to form the confidence interval for the identified region. As discussed in Cheng and Small [2006] and Horowitz and Manski [2000], an alternative Horowitz-Manski style bootstrap confidence interval that takes into account the joint distribution of the estimators of the upper and lower bounds can be formed in the following way. First, an approximation to the joint sampling distribution of $(\hat{\Delta}_{d1}^L, \hat{\Delta}_{d1}^U)$ is obtained through repeated bootstrap sampling. This approximation can then be used to find a critical value $z_n^*$ such that $P(\hat{\Delta}_{d1}^L - z_n^*, \hat{\Delta}_{d1}^U + z_n^*) = 1-\alpha$, where the probability is assessed with respect to the bootstrap distribution. The $100 \times (1-\alpha)\%$ confidence interval for the identified region is then given by $[\hat{\Delta}_{d1}^L - z_n^*, \hat{\Delta}_{d1}^U + z_n^*]$.

As an alternative to inference based on the bootstrap, an approach based on the large sample distributions of the estimators of the individual components of $\hat{\Delta}_{01}^L, \hat{\Delta}_{01}^U, \hat{\Delta}_{21}^L$ and $\hat{\Delta}_{21}^U$ can be used. That is, we can consider the asymptotic behavior of alternative estimators of the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ that we obtain by removing the indicator functions that appear in the expressions for the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ given in the proof of Theorem 1. For $d \in \{0, 2\}$, we let $(\hat{\Gamma}_{d1}^L, \hat{\Gamma}_{d1}^U)$ represent the alternative bounds estimators under Assumptions 1-4 and $(\hat{\Gamma}_{d1}^{L,ER}, \hat{\Gamma}_{d1}^{U,ER})$ represent the alternative bounds estimators under Assumptions 1-5. These alternative bounds estimators will be in agreement with the linear programming-based bounds estimators when the conditions in the indicator functions are satisfied. If these conditions are not satisfied, then the estimators still give valid upper and lower bounds for their respective causal estimand; they simply will not be as tight as the linear programming-based bounds. By Lemma 3, the estimators of the point identified causal outcome parameters as well as the unrestricted upper and lower bounds on the partially identified causal outcome parameters are all $\sqrt{n}$ consistent and asymptotically normal. The alternative bounds estimators just discussed are therefore continuous functions of asymptotically normal estimators and

so by a delta method argument it follows that for $d \in \{0, 2\}$ the bounds estimators $(\hat{\Gamma}^L_{d1}, \hat{\Gamma}^U_{d1})^T$ under Assumptions 1-4 are asymptotically normal, and similarly the bounds estimators $(\hat{\Gamma}^{L,ER}_{d1}, \hat{\Gamma}^{U,ER}_{d1})^T$ under Assumptions 1-5 are asymptotically normal. Furthermore, the asymptotic variances have the general form $\sigma^2 = \nabla h(\boldsymbol{q})\Sigma_q \nabla h(\boldsymbol{q})^T$ for an appropriate choice of function $h$, where $\nabla h(\boldsymbol{q})$ represents the (row vector) gradient of the function $h$ evaluated at $\boldsymbol{q}$. The choice of function $h$ will depend on the set of assumptions made in the analysis, the causal estimand of interest as well as whether it is the upper or lower bound being estimated. In Appendix A Section A.2, we give detailed expressions for these functions. The expressions for the gradients of these functions are notationally cumbersome and so we do not present them. However, we note that the derivatives involved are straightforward and easily computed by a computer algebra system. Given an estimate $\hat{\boldsymbol{q}}$ of $\boldsymbol{q}$, the asymptotic variances can be estimated as $\hat{\sigma}^2 = \nabla h(\hat{\boldsymbol{q}})\hat{\Sigma}_q \nabla h(\hat{\boldsymbol{q}})^T$. An asymptotically valid $100 \times (1-\alpha)\%$ confidence interval for the identified region can then be constructed as $\left[\hat{\Gamma}^L_{d1} - z_{1-\alpha/2} \times \hat{\sigma}^L_{d1}/\sqrt{n},\ \hat{\Gamma}^U_{d1} + z_{1-\alpha/2} \times \hat{\sigma}^U_{d1}/\sqrt{n}\right]$ where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the standard normal distribution.

### *1.7 Checking the Model Assumptions*

Assumptions about the underlying casual model play a key role in any causal analysis. Assumptions often provide analytical advantages, and thus it is important that the set of assumptions used in the analysis be justifiable. While the sample data cannot be used to prove that a given set of assumptions holds, in many instances they can be used to check whether the observed data are inconsistent with certain assumptions. In the instrumental variable literature, the question of empirical tests of the exclusion restriction assumption has received much attention. In the case of a binary outcome and binary instrument, it has been shown [Pearl, 2000] [Bonet, 2013] that the exclusion restriction assumption implies a set of inequalities that the observed data distribution must satisfy in order to be consistent with this assumption. Thus, the observed data can be used to potentially falsify the exclusion restriction assumption. Similarly, in our framework the self-motivated treatment assumption and the exclusion restriction assumption together impose restrictions on the set of permissible observed-data distributions, which can be used to test whether the sample data at hand are in violation of these model assumptions. Under the self-motivated treatment

assumption and the exclusion restriction,

$$\begin{aligned}
q_{01\cdot 0} - q_{01\cdot 1} &= p_{00\cdot 0}\omega_{00} + p_{01\cdot 0}\omega_{01} - p_{00\cdot 1}\omega_{00} \\
&= p_{01\cdot 0}\omega_{01} \\
&\geq 0 \\
q_{11\cdot 1} - q_{11\cdot 0} &= p_{01\cdot 1}\omega_{01} + p_{21\cdot 1}\omega_{21} \\
&\geq 0 \\
q_{00\cdot 0} - q_{00\cdot 1} &= (1 - p_{00\cdot 0})\omega_{00} + (1 - p_{01\cdot 0})\omega_{01} - (1 - p_{00\cdot 1})\omega_{00} \\
&= (1 - p_{01\cdot 0})\omega_{01} \\
&\geq 0 \\
q_{10\cdot 1} - q_{10\cdot 0} &= (1 - p_{01\cdot 1})\omega_{01} + (1 - p_{11\cdot 1})\omega_{11} + (1 - p_{21\cdot 1})\omega_{21} - (1 - p_{11\cdot 0})\omega_{11} \\
&= (1 - p_{01\cdot 1})\omega_{01} + (1 - p_{21\cdot 1})\omega_{21} \\
&\geq 0
\end{aligned}$$

These restrictions provide a way to check whether the observed data violate the model assumptions, by checking whether the sample quantities $\hat{\boldsymbol{q}}$ satisfy

$$\begin{aligned}
\hat{q}_{01\cdot 0} - \hat{q}_{01\cdot 1} &\geq 0 \\
\hat{q}_{11\cdot 1} - \hat{q}_{11\cdot 0} &\geq 0 \\
\hat{q}_{00\cdot 0} - \hat{q}_{00\cdot 1} &\geq 0 \\
\hat{q}_{10\cdot 1} - \hat{q}_{10\cdot 0} &\geq 0.
\end{aligned}$$

If all four of these inequalities are satisfied, then we can conclude that the self-motivated treatment and exclusion restriction assumptions are not contradicted by the sample data. On the other hand, if any of the inequalities are violated, then we wish to know whether the fact that the sample data violate the inequalities indicates that the true observed data distribution violates the inequalities as well. Tests of this hypothesis can be based on the large sample distribution of $\hat{\boldsymbol{q}}$.

### *1.8 Simulations*

In this section we give results from simulations performed under different settings for the underlying causal model parameters. There are several things which we wish to demonstrate with these simulations. First, by considering a range of different scenarios for the underlying causal model parameter values, we wish to show how the bounds we have derived can be highly informative in certain scenarios. Second, we wish to examine the extent to which the consistency of the bounds estimators and the coverage probabilities of the confidence intervals for the identified regions vary with respect to the degree of identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$. In the simulations presented here, we focus on the estimators for upper and lower bounds on the causal estimand $\Delta_{01}$ under Assumptions 1-5. An analogous simulation study for estimators of upper and lower bounds on $\Delta_{21}$ gave results similar to the ones we present below.

To assess the consistency of the estimators, we generated datasets at sample sizes of $n = 100, 200, \ldots, 20000$. Data were generated under three different parameter settings corresponding to three different identifiability scenarios for the causal estimand $\Delta_{01}$: asymmetrically partially identified with uninformative bounds, partially identified with uninformative bounds, and partially identified with informative bounds. These three scenarios correspond to differing levels of information about the causal parameter $p_{01\cdot1}$, which makes up the unidentified portion of the causal estimand $\Delta_{01}$. This was accomplished by varying the stratum membership parameters $\omega_{d_0 d_1}$. For all informativeness scenarios, the potential outcome probabilities in the treatment arm were set to $(p_{00\cdot1}, p_{01\cdot1}, p_{11\cdot1}, p_{21\cdot1}) = (0.03, 0.3, 0.2, 0.2)$ and the potential outcome probabilities in the control arm were set to $(p_{00\cdot0}, p_{01\cdot0}, p_{11\cdot0}, p_{21\cdot0}) = (0.03, 0.03, 0.2, 0.15)$. For the fully unidentified scenario, the stratum membership parameter vector was set to $(\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21}) = (0.05, 0.19, 0.05, 0.71)$, resulting in lower and upper bounds on $p_{01\cdot1}$ of 0 and 1, respectively. For the asymmetrically partially identified scenario, the stratum membership parameter vector was set to $(\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21}) = (0.05, 0.5, 0.05, 0.4)$, resulting in lower and upper bounds on $p_{01\cdot1}$ of 0 and .46, respectively. For the fully informative scenario, the stratum membership parameter vector was set to $(\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21}) = (0.05, 0.85, 0.05, 0.05)$, resulting in lower and upper bounds on $p_{01\cdot1}$ of 0.25 and 0.31, respectively. For each combination of principal

strata distribution $\boldsymbol{\omega}$ and sample size $n$, potential outcomes $(Y_i(0), Y_i(1))$ and potential treatment types $(D_i(0), D_i(1))$ were generated for all $n$ individuals based on the parameter values given above. The observed data were then obtained by randomly assigning $n/2$ individuals to the $Z = 1$ arm and $n/2$ individuals to the $Z = 0$ arm. The observed data were then used to estimate the upper and lower bounds on $\Delta_{01}$ under Assumptions 1-5.

The estimators $\hat{\Delta}_{01}^{U,ER}$ and $\hat{\Delta}_{01}^{L,ER}$ converged to the true values for $\Delta_{01}^{U}$ and $\Delta_{01}^{L}$ in all three data-generating scenarios (Figure 1.3), indicating that the consistency of the bounds estimators is not affected by the degree of identifiability of the causaul estimands. In addition, it is clear from Panel C of Figure 1.3 that the upper and lower bounds for $\Delta_{01}$ can be quite narrow when the principal strata parameter $\omega_{01}$ is sufficiently large.

To assess the performance of the confidence intervals described in Section 1.6, we performed simulations at sample sizes of $n = 1000$ and $n = 10000$. To investigate the effect of the informativeness of the bounds on the coverage probabilities of the confidence intervals, we varied the principal strata probability $\omega_{01}$ from 0.05 to 0.85 by increments of 0.05. As discussed in the description of the consistency simulations, as $\omega_{01}$ increases from 0.05 to 0.85, the causal outcome parameter $p_{01\cdot1}$ goes from fully unidentified to asymmetrically partially identified to partially identified. Given a value for $\omega_{01}$, the principal strata probability $\omega_{21}$ was then set to $\omega_{21} = 0.9 - \omega_{01}$. The principal strata probabilities $\omega_{11}$ and $\omega_{00}$ were both set to 0.05. The potential outcome probabilities for $Y(0)$ and $Y(1)$ were the same as those used in the consistency simulations. For each combination of sample size and data-generating mechanism, we generated 1000 data sets corresponding to the randomized experiment setup described previously. For each data set, we estimated upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ using the linear programming-based bounds. We then constructed two types of 95% confidence intervals for the estimated identified regions. The first type was constructed using the nonparametric bootstrap, while the second type was constructed using the asymptotic distributions of the bounds estimators.

Figure 1.5 shows the estimated coverage probabilities for both types of confidence intervals at both sample size settings. For values of $\omega_{01}$ above approximately 0.70, the coverage probabilities are approximately 95% for both the bootstrap- and asymptotics-based confidence intervals at both sample sizes, indicating that both

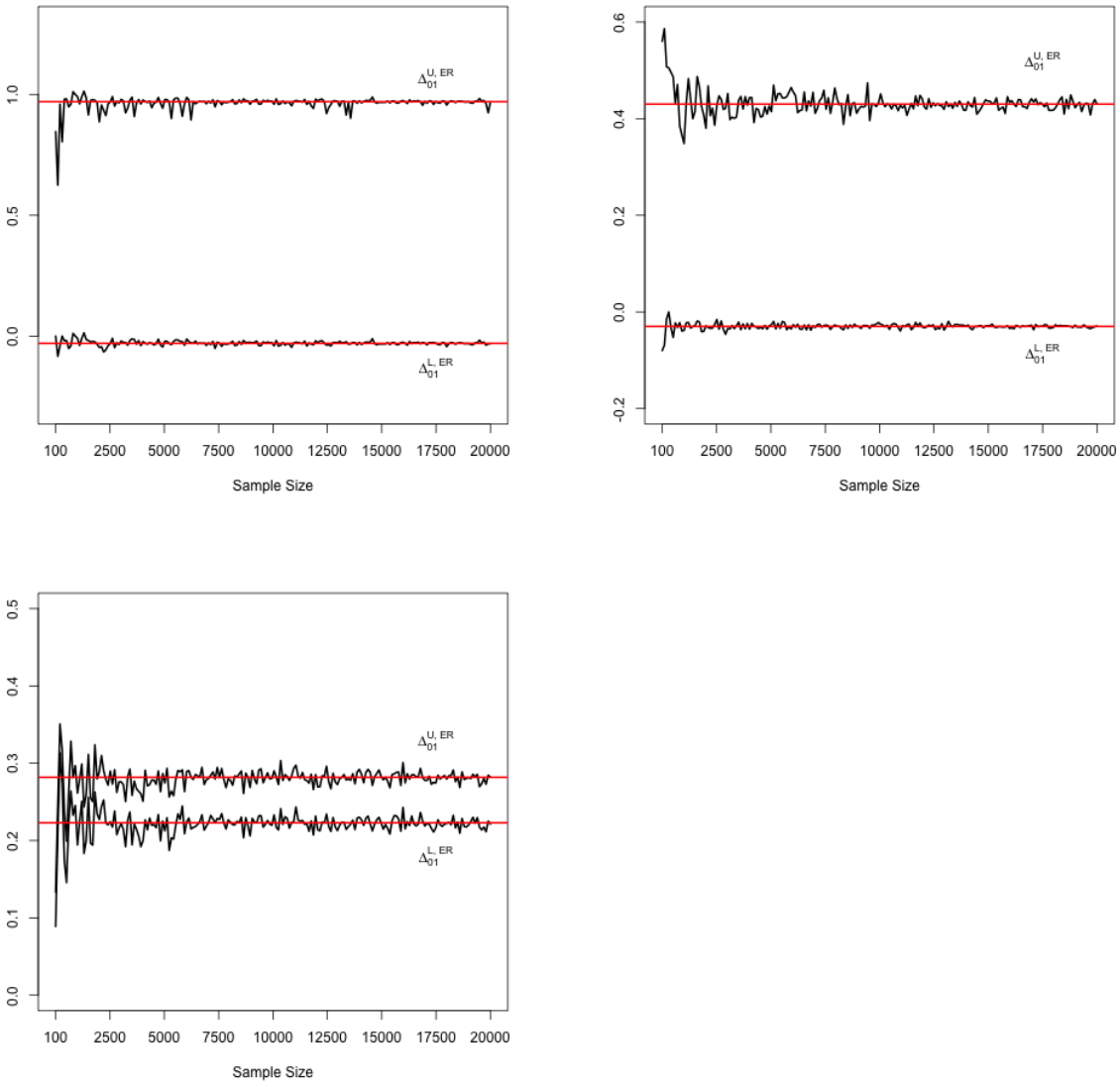
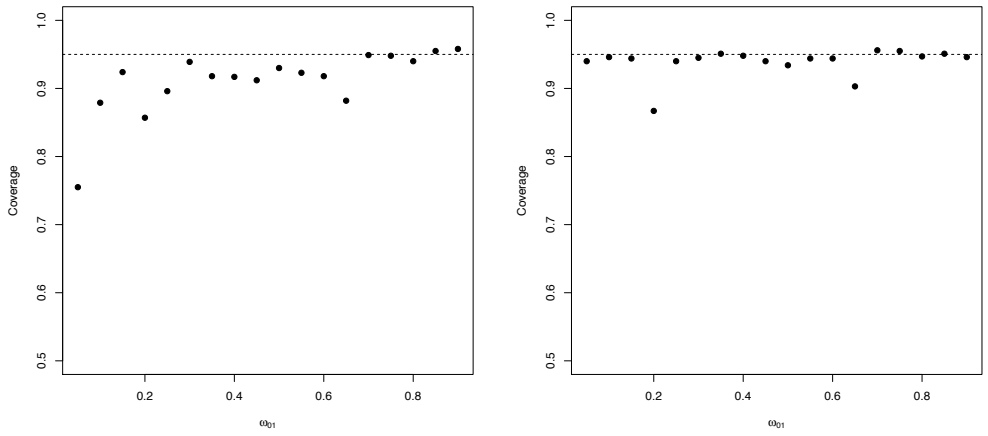
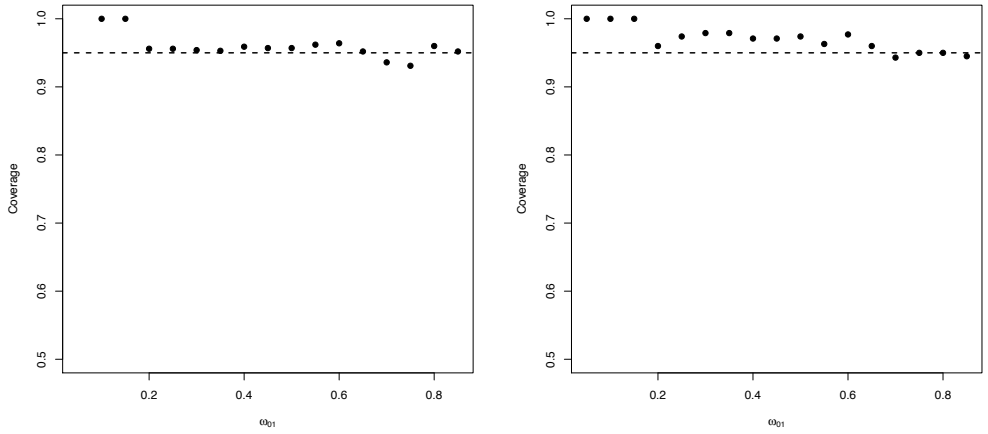



Figure 1.3: Trace plots showing consistency of estimates of $\Delta_{01}^{U,ER}$ and $\Delta_{01}^{L,ER}$ under several identifiability scenarios for the causal parameter $p_{01\cdot 1}$. In the upper lefthand panel, the causal estimand $\Delta_{01}$ is asymmetrically partially identified with uninformative bounds. In the upper righthand panel, $\Delta_{01}$ is partially identified, but the bounds are uninformative. In the lower lefthand panel, $p_{01\cdot 1}$ is partially identified, with informative lower and upper bounds. In each panel, the red (solid) horizontal lines are the true values for $\Delta_{01}^{U,ER}$ and $\Delta_{01}^{L,ER}$




Panel A: Bootstrap confidence intervals for linear programming-based bounds on $\Delta_{01}$.




Panel B: Asymptotic confidence intervals for unrestricted bounds on $\Delta_{01}$.

Figure 1.5: Estimated coverage probabilities for 95% confidence intervals as a function of the parameter $\omega_{01}$. Panel A shows coverages for the bootstrap confidence intervals for the linear programming-based bounds. Panel B shows coverages for the confidence intervals based on asymptotic distributions. In both panels, the figure on the left shows estimated coverages at $n = 1000$, while the figure on the right shows estimated coverages at $n = 10000$.

approaches perform as expected once a sufficient amount of information about the unidentified causal estimand $\Delta_{01}$ (as reflected in the proportion $\omega_{01}$ of the population belonging to the $(D(0) = 0, D(1) = 1)$ principal stratum) is contained in the observed data. However, for values of $\omega_{01}$ below 0.70, we can see that the behavior of the bootstrap- and asymptotics-based confidence intervals differs. The asymptotics-based confidence intervals are conservative, with coverages greater than 95%. On the other hand, the bootstrap-based confidence intervals are anti-conservative, with coverages below 95%. Furthermore, at smaller sample sizes the extent of improper coverage clearly becomes worse as the value of $\omega_{01}$, and hence the relative degree of information about $\Delta_{01}$, decreases. This issue appears to disappear asymptotically, as evidenced by the coverage rates of the bootstrap-based confidence intervals when $n = 10000$, which exhibit coverages of at least 95% fairly uniformly across all values of $\omega_{01}$. These simulations indicate that when the estimated proportion of individuals in the $(D(0) = 0, D(1) = 1)$ principal stratum is relatively low, bootstrap-based confidence intervals are likely to not have correct coverage probabilities. At the same time, when $\hat{\omega}_{01}$ is estimated to be small the resulting bounds on $\Delta_{01}$ are likely to be quite large anyway, making the question of inference for these bounds of relatively little interest.

## *1.9 Application to Early Start Denver Model RCT*

In this section we demonstrate our proposed method using data from an RCT of an early intensive behavioral intervention for children with autism [Dawson et al., 2010]. This RCT was conducted at the University of Washington over a period of two years. The original study found evidence for a beneficial $ITT$ effect of the Early Start Denver Model behavioral intervention when compared to a community-based care control group. As noted in Dawson et al. [2010], the city of Seattle, in which the study was conducted, has well-established community-based treatment options for children with ASD. Thus, the generalizability of the estimated $ITT$ effect found in this study to other populations of children with ASD may be affected by the heterogeneity of treatment received by children in the control group. To aid in understanding how the results from this study might generalize to similar populations with differing availability of community-based care treatment options, we re-analyze the data from the ESDM study using the methodology we have developed for addressing the presence

of multiple versions of control due to a community-based care control condition.

### *1.9.1 Description of the sample*

Data were available for analysis on a total of 45 children with a diagnosis of autism spectrum disorder (ASD) or Pervasive Developmental Disorder-not otherwise specified (PDD-NOS). Children were randomized to either an active treatment group ($n = 23$) or an assess-and-monitor (A/M) group ($n = 21$). Children in the active treatment group were assigned to receive 20 hours per week of the Early Start Denver Model, a naturalistic developmental behavioral intervention for children with ASD. Children in the A/M group did not receive any treatment from the researchers, but did receive information about and referrals to treatment options in the Seattle area. Outcome data were collected at baseline, one year post-randomization, and two years post-randomization. Data on the type of treatment received were collected on a weekly basis. For this analysis, we focus on estimating bounds on $\Delta_{01}$, a comparison between individuals who would have received the ESDM intervention under assignment to treatment and no intervention under assignment to control. As mentioned earlier, such an effect would be of most interest to researchers in an area without well-established treatment options for ASD.

For this analysis, the outcome of interest was an increase in the Mullen Composite score, a measure of cognitive development, of at least 15 points at one year post-randomization compared to baseline. First, we performed the standard Intent-to-Treat analysis, comparing the average outcome in the ESDM intervention group to the average outcome in the community-based control group, without consideration of the type of community-based treatment received by this group. We then analyzed the data using our proposed methodology to take into account both the varying treatment levels in the ESDM group due to non-compliance and the variability in the type of treatment received by those in the control group. For this analysis, we categorized the type of treatment received as follows. Individuals assigned to the ESDM intervention group ($Z = 1$) were categorized as having received the intervention ($D = 1$) if they received an average of 10 hours or more per week of the ESDM intervention during the intervention period. Individuals assigned to the ESDM intervention group who received fewer than 10 hours per week on average of ESDM intervention were categorized as having not received intervention ($D = 0$). For individuals assigned to

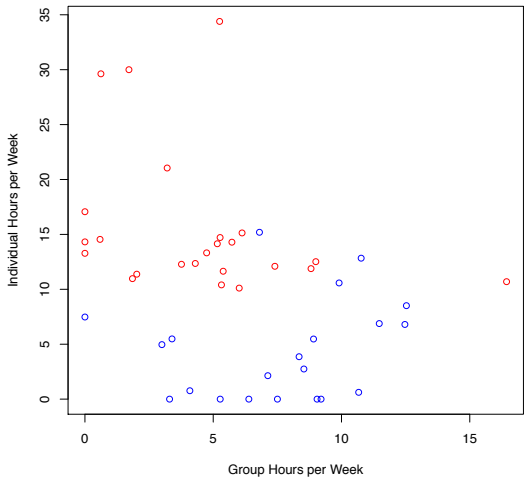



Figure 1.6: Scatterplot of individual versus group average treatment hours at one year post-randomization. The ESDM treatment group is shown in red, while the TAU control group is shown in blue.

the community-based care control group ($Z = 0$), the level of individual treatment was based on the number of hours of individual intensive applied behavioral analysis (ABA) therapy. Children who received greater than 10 hours per week on average of individualized treatment were categorized as having received individual-based treatment ($D = 1$). Children who received fewer than 10 hours per week on average of individualized treatment and greater than 10 hours per week of group treatment were categorized as having received group-based treatment ($D = 2$). Children who received fewer than 10 hours per week of individualized treatment and fewer than 10 hours per week of group treatment were categorized as having not received treatment ($D = 0$).

Among those assigned to the ESDM intervention ($Z = 1$), the proportion with an increase in Mullen Composite Score of at least 15 points at one-year post-randomization compared to baseline was 0.54. Among those assigned to the A/M control group ($Z = 0$), the proportion with an increase in Mullen Composite Score of at least 15 points at one-year post-randomization compared to baseline was 0.14. The estimated $ITT$ effect was 0.40 (95% CI: 0.14, 0.66).

We applied the model checks to the observed-data and found that they were not in violation of Assumptions 1-5, so we will base our analysis on those assumptions. The estimated principal strata probabilities were $(\hat{\omega}_{00}, \hat{\omega}_{01}, \hat{\omega}_{11}, \hat{\omega}_{21}) = (0, 0.67, 0.19, 0.14)$. The identified region for the $\Delta_{01}$ was $(\hat{\Delta}_{01}^{L}, \hat{\Delta}_{01}^{U}) = (0.31, 0.60)$, with a 95% bootstrap confidence interval of $(-0.13, 0.94)$ and an asymptotic 95% confidence interval of $(-0.10, 0.99)$.

| Estimand | Estimate | 95% Confidence Interval |
|---|---|---|
| ITT | 0.4 | (0.14, 0.66) |
| $\Delta_{01}$ (with asymptotic-based CI) | (0.31, 0.60) | (-.10, 0.99) |
| $\Delta_{01}$ (with bootstrap-based CI) | (0.31, 0.60) | (-0.13, 0.94) |

Table 1.3: Point and interval estimates for the $ITT$ effect and $\Delta_{01}$ at 1 year post-randomization.

The analysis of the baseline to 1 year post-randomization data shows how the methods we have developed can be applied to real data sets to go beyond the standard $ITT$ analysis when analyzing data from RCTs with multiple versions of control. The estimated upper and lower bounds on $\Delta_{01}$ resulted in an estimate of the identified region that was both informative and relatively tight, showing that our methods can provide a good deal of additional information about causal estimands of interest despite the lack of point identifiability. At the same time, the width of the bootstrap and asymptotic confidence intervals for the estimated identified region shows how the additional uncertainty around partially identified parameters can affect precise inference once sampling variability is taken into account.

## *1.10 Discussion*

Many interventions of interest lack a feasible placebo version that can be used as a control condition in an RCT. Behavioral interventions are an important class of such interventions. We have developed a framework for causal inference in RCTs with control conditions based on the community-care or treatment-as-usual model often used in RCTs of behavioral interventions. In a randomized trial using a community-based care control condition, the distribution of the community-based treatment options in

the sample will be representative of the distribution of community-based treatment options in the population. Thus, even in the presence of control group heterogeneity, the $ITT$ effect provides valuable information about the effect of the intervention in the community in which the study was performed. Additionally, the $ITT$ effect can be generalized to other communities with the same distribution of community-based treatment options as the community in which the study was performed. Issues arise, however, when trying to generalize to a population with a different availability of community-based treatment choices. By looking at stratum-specific effects, the approach presented in this paper can aid in understanding the generalizability of results from randomized trials with control group heterogeneity. This is especially important in the case of behavioral interventions since RCTs of these interventions are typically conducted in settings where control group heterogeneity under the community-based care protocol is likely to be present, such as cities with established community-based treatment options or areas near major research universities. For researchers in a community with very few community-based treatment options, the causal estimand $\Delta_{01}$ is arguably more appropriate when interpreting the results of a study done in a population with well-established community-based treatment options. Conversely, the causal estimand $\Delta_{21}$ may be more appropriate for researchers in a community with well-established treatment options that differ from the intervention being studied. In this article, we have demonstrated the identifiability issues that are present when data from a single population are available. We have derived bounds on these partially identified causal estimands of interest expressible in terms of identified quantities. The bounds can then be estimated using sample equivalents to the observed-data quantities and inference can be based on large sample approximations or resampling methods such as the bootstrap.

By focusing on conditional causal effects in a principal stratification framework, our approach differs from previous approaches that have focused on estimating marginal effects through covariate adjustment. In theory, one could use these adjustment-based approaches in our setting as well to estimate the marginal causal effects of receiving the intervention compared to receiving each of the possible versions of control available in the community. However, this approach relies on having data on all confounding variables, thus eliminating one of the key advantages of a randomized design. The approach we have proposed takes advantage of the randomized design and only requires

that data be available on the versions of treatment received. One of the drawbacks of focusing on principal strata effects, of course, is that one cannot know which principal stratum a particular individual belongs to, as principal stratum membership is an unobserved latent characteristic. As a result, the question of how to use estimates of principal causal effects to inform policy is a complicated one. We have motivated our examination of RCTs with multiple versions of control with the example of a researcher or practitioner trying to decide whether the results of an RCT conducted in another geographic using a TAU control can be generalized to the population that he or she works with. The availability of community-based care in this researcher's geographic area may be very limited, or there may be an established form of treatment being used in typical practice that is very different from the intervention being studied. In the former case, the researcher may find $\Delta_{01}$ to be of primary interest, since he or she can reasonably infer that most potential recipients of the intervention would not receive any form of treatment otherwise due to the lack of community-based care options in his or her geographic area. In the latter case, $\Delta_{21}$ could be of interest, since the researcher may infer that most individuals would receive an alternative form of care if not given the intervention. Thus, we see causal effects such as $\Delta_{01}$ and $\Delta_{21}$ as having potential for practical use as a way of understanding how the intervention effects observed in one area may generalize to other areas.

Our approach can also be used in situations where the exclusion restriction is either unlikely to hold theoretically or is directly contradicted by the data. Obtaining results that do not rely on the exclusion restriction is especially important for analyzing results from RCTs of behavioral interventions that use the TAU control condition, since the lack of a placebo means that blinding participants and investigators to treatment assignment may not be possible.

Although we have focused on frequentist inference, Bayesian inference for this problem could be performed in several ways. Bayesian inference for the point identified bounds could be conducted by specifying a prior distribution for the observed data quantities, similar to the approach used in Richardson et al. [2011]. Alternatively, the partial identification framework for Bayesian analysis outlined in Gustafson [2015] could be applied.

When constructing confidence intervals in a partial identification setting, the analyst must choose between intervals for the entire identified region and intervals for

the unidentified parameter. In Imbens and Manski [2004], the authors give a method for constructing the latter type of intervals. However, their approach relies on the underlying model satisfying certain regularity conditions that do not hold for the model given here. We have therefore focused on the entire identified region as our target for inference. Finding a way to construct intervals for the partially identified estimands $\Delta_{01}$ and $\Delta_{21}$ themselves, possibly by modifying the approach given in Imbens and Manski [2004], is an area for future research.

Another area for future research is the extension of this framework from a three category representation for the control group heterogeneity to a more general $k$ category representation. We have focused on the three category representation (in which those assigned to the control group are categorized into either not receiving any treatment, receiving treatment similar to the intervention, or receiving treatment distinct from the intervention) based on observed patterns of control group heterogeneity in our motivating example of RCTs of behavioral interventions for autism. Finer classifications of control group heterogeneity may be more appropriate for other contexts. Extending the model described here to include a finer representation of control group heterogeneity would involve increasing the number of principal strata. However, as formulated here, the width of the identified region for the causal effect within a principal stratum tends to increase as the proportion of the population that belongs to that principal stratum decreases. Thus, a straightforward extension of the model described here will likely lead to wide bounds on the within-strata causal effects. One way to address this issue would be to incorporate available covariate information in order to give sharper bounds.

A common criticism of approaches based on partial identifiability is that they are not as useful for decision making due to the additional uncertainty about the parameters of interest. However, in situations where the data are limited, approaches based on partial identifiability may be the only available options that do not rely on imposing additional, potentially unrealistic assumptions. Methods such as the ones proposed here offer a way to gain information from otherwise limited data with a minimal number of assumptions. Although not as definitive as providing a single point estimate, placing bounds on parameters of interest can still be used for many of the same purposes, such as determining the relative magnitude of an effect or addressing the question of whether an effect exists. At the very least, researchers

should be aware of the generalizability issue introduced by the use of TAU control conditions. At the same time, authors of studies using TAU control conditions can assist other researchers and decision makers in interpreting the results by providing detailed information about the distribution of community-based treatment options available in the population in which their study was conducted, in addition to the standard sample demographic characteristics.

# Chapter 2

# CAUSAL INFERENCE IN RCTS WITH MULTIPLE VERSIONS OF CONTROL WITH DATA FROM MULTIPLE POPULATIONS

## *2.1 Introduction*

In this chapter, we consider an extension of the framework developed in Chapter 1 when data from multiple populations are available. The extension we consider is based on the idea of a multi-site RCT, in which the same intervention is implemented in two or more similar but geographically distinct populations. Multi-site trials are often used as a way to increase sample sizes [Friedman et al., 2015], an especially important issue for behavioral therapy interventions [Howlin and Magiati, 2009]. In this chapter, we examine the impact that the natural variability between sites in the distribution of community-based care options can have on the identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$ that are unidentified when data from only a single population are available. Identifiability is fundamentally an issue of a lack of sufficient information in the data. Often this manifests as a discrepancy between the number of model parameters and the number of functionally independent observed-data quantities. One common approach to obtaining identifiability in situations where the data do not provide enough information is to make additional assumptions that reduce the number of parameters. A classic example of this "model contraction" approach is the exclusion restriction assumption, which reduces the number of parameters necessary for characterizing the underlying causal model. The exclusion restriction plays a vital role in identifying the Complier Average Causal Effect (CACE) in the binary non-compliance framework (see Yau and Little [2001], Zhou and Li [2006], and Taylor and Zhou [2009] for examples of the exclusion restriction used in different contexts to identify the CACE, as well as Imbens and Rubin [1997] and Hirano et al. [2000] for discussion of the effect of relaxing the exclusion restriction).

The drawback of the model contraction approach is that the assumptions neces-

sary for obtaining identifiability may not be reasonable. And, as seen in the previous chapter, even fairly strong assumptions such as the exclusion restriction assumption and the self-motivated treatment assumption are sometimes not sufficient to obtain identifiability. As discussed in Gustafson [2005b], an alternative, somewhat counter-intuitive approach to obtaining identifiability is to actually expand the model to include more parameters. This approach has been used in the literature on diagnostic testing as a way to address the known identifiability issues for sensitivity and specificity parameters in the absence of a gold standard [Hui and Walter, 1980] [Jones et al., 2010]. As noted in Jones et al. [2010], the sensitivity and specificity of an imperfect diagnostic test are not point identified in the absence of a gold standard test when data are only available from a single population. To solve these identifiability issues, the authors consider an approach based on applying multiple imperfect tests to multiple populations. While this expands the model by introducing new parameters, it also introduces additional observed-data quantities. Importantly, their model expansion approach assumes that certain parameters are equal across populations. Recently in the causal inference literature, Jiang et al. [2016] used a model expansion-type approach similar to the one we consider here to obtain identifiability of principal causal effects in the setting of an RCT with a binary surrogate outcome measured post-randomization. As in Jones et al. [2010], the approach used in Jiang et al. [2016] required additional assumptions about the parameters introduced in the expanded model. Thus, both the model contraction approach and the model expansion approach generally require additional assumptions above and beyond those made for the original unidentified model. The key difference between the two approaches is that the model expansion approach makes additional assumptions due to the addition of extra sources of data, while the model contraction approach uses additional assumptions in the place of additional data.

We will begin by establishing general notation, assumptions and preliminary results that will be used in all of the scenarios we consider in this chapter. We then examine several scenarios of interest. The first scenario is the case of a two-site trial where the exclusion restriction can be assumed to hold. The second scenario is the generalization of the results obtained in the two-site case to the general case of a trial with two or more sites. The third scenario is the case of a two-site trial where the exclusion restriction assumption is not made. The fourth scenario is the generalization

of the results obtained in the two-site case to the general case of a trial with two or more sites.

As in Chapter 1, our motivating example will be an RCT of a behavioral intervention using a TAU control group due to the lack of a plausible placebo, which leads to individuals in the control group receiving one of several versions of the control condition post-randomization. We will again assume that the versions of control can be categorized as either 1.) receiving no treatment, 2.) receiving a form of active treatment equivalent to the intervention being studied, or 3.) receiving a form of active treatment distinct from the intervention being studied.

## *2.2 Notation*

We assume that we have data available from a multi-site RCT of an intervention of interest subject to control group heterogeneity. The observations take the form of quadruples $(Y_i, D_i, Z_i, S_i)$. We will assume that observations are independent both within and across sites. The random variables $Y_i, D_i$, and $Z_i$ are defined in the same way as in Chapter 1. The variable $S_i$ is a categorical variable indicating the site that the $i^{th}$ individual belongs to. Thus, $S_i$ takes values in $\{1, 2, \ldots, K\}$ where $K$ is the number of sites participating in the RCT. To simplify notation, we will assume that the treatment assignment probability is the same across sites, $P(Z_i = 1 \mid S_i = k) = p_z$. Site-specific randomization probabilities can be easily accounted for as well. We will let $n_k, k \in \{1, 2, \ldots, K\}$ be the sample size for the $k^{th}$ site, and $n = \sum_1^K n_k$ be the total sample size.

We will parameterize the causal model in terms of the causal parameters conditional on site. We will let $\omega_{d_0 d_1 \cdot k} = P(D(0) = d_0, D(1) = d_1 \mid S = k)$ represent the probability of belonging to the principal stratum $(D(0) = d_0, D(1) = d_1)$ for individuals at site $k$. We will let $p_{d_0 d_1 \cdot zk} = P(Y(z) = 1 \mid D(0) = d_0, D(1) = d_1, S = k)$ represent the conditional probability of success (as defined by the intervention of interest) for potential outcome $Y(z)$, where we condition on both the principal stratum $(D(0) = d_0, D(1) = d_1)$ and the site $S = k$. Similar to Chapter 1, we will let $q_{dy \cdot zk} = P(Y = y, D = d \mid Z = z, S = k)$ designate the observed-data quantities in the $k^{th}$ site.

For the $k^{th}$ site, we will let

$$\boldsymbol{\theta}_k = (p_{00\cdot 0k}, p_{01\cdot 0k}, p_{11\cdot 0k}, p_{21\cdot 0k}, p_{00\cdot 1k}, p_{01\cdot 1k}, p_{11\cdot 1k}, p_{21\cdot 1k}, \omega_{00\cdot k}, \omega_{01\cdot k}, \omega_{11\cdot k}, \omega_{21\cdot k})^T$$

designate the vector of causal model parameters and

$$\boldsymbol{q}_k = (q_{01\cdot 0k}, q_{11\cdot 0k}, q_{10\cdot 0k}, q_{20\cdot 0k}, q_{21\cdot 0k}, q_{00\cdot 1k}, q_{01\cdot 1k}, q_{11\cdot 1k})^T$$

designate the vector of observed-data quantities in the $k^{th}$ site implied by the underlying causal model $\boldsymbol{\theta}_k$. We will let $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_K)^T$ represent the combined vector of observed-data quantities from the $K$ sites.

### *2.3 Assumptions*

In this section, we detail the assumptions we will make going forward. The assumptions can be divided into two categories. The first category pertains to assumptions about the causal model that are assumed to hold within each site. The second category pertains to assumptions about the relationship between the sites themselves. We begin by detailing the within-site assumptions. First, we will make the three commonly used assumptions of non-interference, exchangeability due to randomization, and independence and identical distribution of observations. Non-interference states that the counterfactual quantities for the $i^{th}$ individual are not related to the counterfactual quantities for the other individuals. Exchangeability states that the counterfactual values are independent of the treatment assignment, which follows from the assumption that $Z$ is randomized. Finally, the independence and identical distribution assumption states that the joint distribution of the observed data for each individual $(Y_i, Z_i, D_i, S_i)$ can be modeled as coming from a common true distribution function $F$.

**Assumption 1** *a.)* Non-interference.
*Let $\boldsymbol{z}$ and $\boldsymbol{d}$ be vectors of treatment assignments and treatment versions for the $n$ individuals in the sample. Then for all $i = 1, 2, \ldots, n$,*

$$Y_i(\boldsymbol{z}, \boldsymbol{d}) = Y_i(z_i, d_i) \text{ and } D_i(\boldsymbol{z}) = D_i(z_i).$$

*b.)* Exchangeability/randomization.
*For all* $i = 1, 2, \ldots, n, z \in \{0, 1\}$, *and* $d \in \mathcal{D}_z$

$$Z_i \perp\!\!\!\perp D_i(z), Y_i(z, d).$$

*c.)* Independent, identically distributed (iid) data.
*For all* $i = 1, 2, \ldots, n$,

$$(Y_i, D_i, Z_i, S_i) \overset{iid}{\sim} F$$

*for some distribution* $F$.

The distribution $F$ of the observed-data, and in particular the associated conditional distributions $F_k$ of the observed-data within each site $k \in \{1, 2, \ldots, K\}$, will be discussed in greater detail later in this chapter. We will also make the following extended consistency assumption, which establishes the connection between the counterfactual values and the observed data.

**Assumption 2** Extended consistency assumption.
*For all* $i$,

$$D_i = D_i(0) \times (1 - Z_i) + D_i(1) \times Z_i$$

*and*

$$\begin{aligned} Y_i = {} & Y_i(0,0) \times (1 - Z_i)\mathbb{1}(D_i = 0) + Y_i(0,1) \times (1 - Z_i)\mathbb{1}(D_i = 1) + Y_i(0,2) \times (1 - Z_i)\mathbb{1}(D_i = 2) \\ & + Y_i(1,0) \times Z_i\mathbb{1}(D_i = 0) + Y_i(1,1) \times Z_i\mathbb{1}(D_i = 1). \end{aligned}$$

This formulation of the consistency assumption emphasizes how the potential outcome that is observed through the sampling process is only known to us so long as we have data on the version of the assigned treatment actually received. If this information is missing, then we cannot determine which potential outcome we have observed. For instance, suppose that for the $i^{th}$ individual we observe $Z_i = 0$ and $Y_i = 1$. Given this information alone, we cannot link the observed data for the $i^{th}$ individual to his

or her potential outcome because we cannot determine whether we have observed $Y_i(0,0), Y_i(0,1)$, or $Y_i(0,2)$.

We will make a positivity assumption for the principal strata probabilities $\omega_{01\cdot k}$ and $\omega_{21\cdot k}$ for each site $k \in \{1, 2, \ldots, K\}$.

**Assumption 3** Positivity assumption.
*For the $k^{th}$ site, $k \in \{1, 2, \ldots, K\}$, the principal strata probabilities $\omega_{01\cdot k}$ and $\omega_{21\cdot k}$ are both non-zero.*

We will also make the following assumption about the potential versions of treatment in the $Z = 0$ and $Z = 1$ arms.

**Assumption 4** Self-motivated treatment assumption.
*For each individual $i$ in the $k^{th}$ site, $k \in \{1, 2, \ldots, K\}$, we have that*

$$D_i(0) \geq 1 \implies D_i(1) = 1.$$

This assumption is based on the idea that the community-based treatment chosen by an individual in the TAU control group corresponds to treatment that the individual must seek out on their own. It is reasonable, then, to assume that if an individual would seek out a type of active treatment on their own, then they would similarly accept a treatment being offered to them as part of a study. We note that this assumption is similar to the commonly invoked monotonicity assumption in the binary non-compliance literature. The self-motivated treatment assumption has the effect of reducing the number of principal strata from six to four. Table 1.1 in Chapter 1 summarizes the principal strata and the corresponding probability parameter for each stratum. Going forward, we will let $\mathcal{D}$ stand for the subset of pairs $(d_0, d_1)$ in the Cartesian product $\mathcal{D}_0 \times \mathcal{D}_1$ that are permissible under the self-motivated treatment assumption.

Finally, we will have reason to consider situations where the following assumption holds.

**Assumption 5** Exclusion restriction assumption.
*For each site* $k \in \{1, 2, \ldots, K\}$*, if* $d_0 = d_1$*, then* $p_{d_0 d_1 \cdot 1k} = p_{d_0 d_1 \cdot 0k} \equiv p_{d_0 d1 \cdot k}$.

This assumption states that the assignment mechanism has no effect on the potential outcomes $Y(0)$ and $Y(1)$ except through the version of treatment actually received. This assumption is related to the instrumental variable analysis often seen in the econometrics literature [Angrist et al., 1996]. When appropriate, it allows for improved estimation and inference. However, it is not necessary for all of the results that follow.

We will additionally make the following assumption about the comparability of the causal parameters between sites

**Assumption 6** Comparable Strata Effects.
*For each pair of sites* $k_1, k_2$ *and for* $d \in \{0, 2\}$ *,* $p_{d1 \cdot 1k_1} - p_{d1 \cdot 0k_1} = p_{d1 \cdot 1k_2} - p_{d1 \cdot 0k_2} = \Delta_{d1}$ *and* $\omega_{d1 \cdot k_1} \neq \omega_{d1 \cdot k_2}$.

In other words, we will assume that the site populations share common within-strata intervention effects for strata $(D(0) = 0, D(1) = 1)$ and $(D(0) = 2, D(1) = 1)$, but differ with respect to the distribution of those strata. This assumption does not require that the causal outcome parameters $p_{dy \cdot zk}$ be equal between sites, but simply that $p_{01 \cdot 1k} - p_{01 \cdot 0k} = p_{01 \cdot 1k'} - p_{01 \cdot 0k'}$ and similarly that $p_{21 \cdot 1k} - p_{21 \cdot 0k} = p_{21 \cdot 1k'} - p_{21 \cdot 0k'}$. This assumption will be crucially important for the identifiability results we obtain. We discuss its implications for the overall model in a later section.

We will also make the following assumption about the behavior of the sample sizes in each of the $K$ sites.

**Assumption 7** Proportional Sample Sizes *The sample sizes across sites are chosen in such a way that for all* $n$

$$P(S_i = k) = p_k \in (0, 1), k \in \{1, 2, \ldots, K\}$$

where $\sum_k p_k = 1$. This assumption ensures that no single site can dominate the other sites in terms of its contribution to the total sample size as we consider the large sample behavior of the estimators we derive in this chapter. As stated, the underlying probability that an individual belongs to a particular site does not depend on the

total sample size $n$. Typically in situations where data from multiple populations are available (e.g. the two sample t-test), this assumption is instead formulated in terms of a statement about the limiting ratio of the sample sizes in each group as the total sample size goes to infinity, thus allowing for the probability of belonging to a particular site to change depending on the total sample size. We could have formulated this assumption in those terms as well, but have chosen to forego this for the sake of simplicity. An extension to allow for changing site proportions would be straightforward.

As stated above, the assumptions made in this chapter can be categorized into within-site assumptions (Assumptions 1- 5) and between-site assumptions (Assumptions 6 -8). As in Chapter 1, we will have reason to consider situations in which the exclusion restriction does and does not hold. Going forward, we will use WS-ER (Within-Site Exclusion Restriction) to refer to the set of within-site assumptions including the exclusion restriction (Assumptions 1-5) and the set of between-site assumptions (Assumptions 6-8) collectively, and WS-NER (Within-Site No Exclusion Restriction) to refer to the set of within-site assumptions not including the exclusion restriction (Assumptions 1-4) and the set of between-site assumptions (Assumptions 6-8) collectively.

### *2.4 Estimation and Inference for Observed-data Quantities*

Similar to Chapter 1, we will parameterize the observed-data distribution within each site in terms of the joint probabilities $q_{dy\cdot zk} = P(Y = y, D = d \mid Z = z, S = k)$. We will let $\boldsymbol{q}_k = (q_{01\cdot 0k}, q_{11\cdot 0k}, q_{10\cdot 0k}, q_{20\cdot 0k}, q_{21\cdot 0k}, q_{00\cdot 1k}, q_{01\cdot 1k}, q_{11\cdot 1k})^T$ represent the vector of observed-data quantities for the $k^{th}$ site, and $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_K)^T$ represent the combined vector of observed-data quantities for all $K$ sites.

The observed-data quantities $q_{dy\cdot zk} = P(Y = y, D = d \mid Z = z, S = k)$ can be estimated by their sample counterparts

$$\hat{q}_{dy\cdot zk} = \frac{\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(D_i = d)\mathbb{1}(Y_i = y)\mathbb{1}(S_i = k)}{\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(S_i = k)}.$$

We will let $\hat{\boldsymbol{q}}_k$ be the vector of estimated observed-data quantities for the $k^{th}$ site, and $\hat{\boldsymbol{q}} = (\hat{\boldsymbol{q}}_1, \hat{\boldsymbol{q}}_2, \ldots, \hat{\boldsymbol{q}}_K)$ be the combined vector of observed-data quantities for all sites. The asymptotic behavior of $\hat{\boldsymbol{q}}$ is given in the following lemma.

**Lemma 6** *Under either Assumptions WS-ER or Assumptions WS-NER,*
*a.)* $\hat{\boldsymbol{q}}$ *is a consistent estimator for* $\boldsymbol{q}$

$$\hat{\boldsymbol{q}} \xrightarrow{p} \boldsymbol{q}$$

*and*
*b.)* $\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q})$ *is asymptotically normal with distribution*

$$\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q}) \xrightarrow{d} N_{8K}(\mathbf{0}, \Sigma_q)$$

*where* $\Sigma_q$ *is the* $8K \times 8K$ *block diagonal matrix*

$$\Sigma_q = \begin{pmatrix} \Sigma_{q1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_{q2} & \dots & \mathbf{0} \\ \vdots & \dots & \ddots & \vdots \\ \mathbf{0} & \vdots & \vdots & \Sigma_{qK} \end{pmatrix}$$

*consisting of the site-specific covariance matrices* $\Sigma_{qk}, k \in \{1, 2, \dots K\}$.

The site-specific covariance matrices $\Sigma_{qk}$ have the form of a block diagonal matrix as well,

$$\Sigma_{qk} = \begin{pmatrix} \frac{1}{p_{0k}}\Sigma_0 & \mathbf{0} \\ \mathbf{0} & \frac{1}{p_{1k}}\Sigma_1 \end{pmatrix}$$

where $p_{zk} = P(Z = z, S = k) = P(Z = z \mid S = k)P(S = k) = p_z p_k$,

$$\Sigma_{0k} = \begin{array}{c} \\ q_{01\cdot 0k} \\ q_{11\cdot 0k} \\ q_{10\cdot 0k} \\ q_{20\cdot 0k} \\ q_{21\cdot 0k} \end{array} \begin{array}{c} \begin{array}{ccccc} q_{01\cdot 0k} & q_{11\cdot 0k} & q_{10\cdot 0k} & q_{20\cdot 0k} & q_{21\cdot 0k} \end{array} \\ \begin{pmatrix} q_{01\cdot 0k}(1-q_{01\cdot 0k}) & -q_{01\cdot 0k}q_{11\cdot 0k} & -q_{01\cdot 0k}q_{10\cdot 0k} & -q_{01\cdot 0k}q_{20\cdot 0k} & -q_{01\cdot 0k}q_{21\cdot 0k} \\ -q_{01\cdot 0k}q_{11\cdot 0k} & q_{11\cdot 0k}(1-q_{11\cdot 0k}) & -q_{10\cdot 0k}q_{11\cdot 0k} & -q_{11\cdot 0k}q_{20\cdot 0k} & -q_{11\cdot 0k}q_{21\cdot 0k} \\ -q_{01\cdot 0k}q_{10\cdot 0k} & -q_{10\cdot 0k}q_{11\cdot 0k} & q_{10\cdot 0k}(1-q_{10\cdot 0k}) & -q_{10\cdot 0k}q_{20\cdot 0k} & -q_{10\cdot 0k}q_{21\cdot 0k} \\ -q_{01\cdot 0k}q_{20\cdot 0k} & -q_{11\cdot 0k}q_{20\cdot 0k} & -q_{10\cdot 0k}q_{20\cdot 0k} & q_{20\cdot 0k}(1-q_{20\cdot 0k}) & -q_{20\cdot 0k}q_{21\cdot 0k} \\ -q_{01\cdot 0k}q_{21\cdot 0k} & -q_{11\cdot 0k}q_{21\cdot 0k} & -q_{10\cdot 0k}q_{21\cdot 0k} & -q_{20\cdot 0k}q_{21_c dot 0k} & q_{21\cdot 0k}(1-q_{21\cdot 0k}) \end{pmatrix} \end{array}$$

and

$$\Sigma_{1k} = \begin{array}{c} \\ q_{00\cdot 1k} \\ q_{01\cdot 1k} \\ q_{11\cdot 1k} \end{array} \begin{array}{c} \begin{array}{ccc} q_{00\cdot 1k} & q_{01\cdot 1k} & q_{11\cdot 1k} \end{array} \\ \begin{pmatrix} q_{00\cdot 1k}(1-q_{00\cdot 1k}) & -q_{00\cdot 1k}q_{01\cdot 1k} & -q_{00\cdot 1k}q_{01\cdot 1k} \\ -q_{00\cdot 1k}q_{01\cdot 1k} & q_{01\cdot 1k}(1-q_{01\cdot 1k}) & -q_{01\cdot 1k}q_{11\cdot 1k} \\ -q_{00\cdot 1k}q_{11\cdot 1k} & -q_{01\cdot 1k}q_{11\cdot 1k} & q_{11\cdot 1k}(1-q_{11\cdot 1k}) \end{pmatrix} \end{array}.$$

The proof of this lemma is given in Appendix B.

## *2.5 Identifiability with Data from Two Populations under the Exclusion Restriction*

We now consider the case of a two-site RCT under Assumptions WS-ER, so that data are available from $k = 2$ sites and the exclusion restriction is assumed to hold within each site. The impact of the Comparable Strata Effects assumption combined with the exclusion restriction on the identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$ can be seen by recalling the following relationship between the observed-data quantities and the unidentified causal model parameters $p_{01\cdot 1}$ and $p_{21\cdot 1}$ from a single-site RCT under the exclusion restriction :

$$q_{11\cdot 1} - q_{11\cdot 0} = p_{01\cdot 1}\omega_{01} + p_{21\cdot 1}\omega_{21}.$$

When data from two sites are available, this relationship holds within each site

$$\begin{aligned} q_{11\cdot 11} - q_{11\cdot 01} &= p_{01\cdot 11}\omega_{01\cdot 1} + p_{21\cdot 11}\omega_{21\cdot 1}, \\ q_{11\cdot 12} - q_{11\cdot 02} &= p_{01\cdot 12}\omega_{01\cdot 2} + p_{21\cdot 12}\omega_{21\cdot 2}. \end{aligned} \tag{2.1}$$

Subtracting $p_{01\cdot 01}\omega_{01\cdot 1} + p_{21\cdot 01}\omega_{21\cdot 1}$ from the first equation in (2.1) yields

$$\begin{aligned} q_{11\cdot 11} - q_{11\cdot 01} - p_{01\cdot 01}\omega_{01\cdot 1} - p_{21\cdot 01}\omega_{21\cdot 1} &= p_{01\cdot 11}\omega_{01\cdot 1} + p_{21\cdot 11}\omega_{21\cdot 1} - p_{01\cdot 0}\omega_{01\cdot 1} - p_{21\cdot 01} \\ &= (p_{01\cdot 11} - p_{01\cdot 01})\omega_{01\cdot 1} + (p_{21\cdot 11} - p_{21\cdot 01})\omega_{21\cdot 1} \\ &= \Delta_{01}\omega_{01\cdot 1} + \Delta_{21}\omega_{21\cdot 1}. \end{aligned}$$

Similarly, subtracting $p_{01\cdot 02}\omega_{01\cdot 2} + p_{21\cdot 02}\omega_{21\cdot 2}$ from the second equation in (2.1)

yields

$$\begin{aligned} q_{11\cdot 12} - q_{11\cdot 02} - p_{01\cdot 02}\omega_{01\cdot 2} - p_{21\cdot 02}\omega_{21\cdot 2} &= p_{01\cdot 12}\omega_{01\cdot 2} + p_{21\cdot 12}\omega_{21\cdot 2} - p_{01\cdot 02}\omega_{01\cdot 2} - p_{21\cdot 02} \\ &= (p_{01\cdot 12} - p_{01\cdot 02})\omega_{01\cdot 2} + (p_{21\cdot 12} - p_{21\cdot 02})\omega_{21\cdot 2} \\ &= \Delta_{01}\omega_{01\cdot 2} + \Delta_{21}\omega_{21\cdot 2}. \end{aligned}$$

This gives the following two equations

$$\begin{aligned} q_{11\cdot 11} - q_{11\cdot 01} - p_{01\cdot 01}\omega_{01\cdot 1} - p_{21\cdot 01}\omega_{21\cdot 1} &= \Delta_{01}\omega_{01\cdot 1} + \Delta_{21}\omega_{21\cdot 1}, \\ q_{11\cdot 12} - q_{11\cdot 02} - p_{01\cdot 02}\omega_{01\cdot 2} - p_{21\cdot 02}\omega_{21\cdot 2} &= \Delta_{01}\omega_{01\cdot 2} + \Delta_{21}\omega_{21\cdot 2}. \end{aligned}$$

The only unidentified quantities in these equations are the causal estimands $\Delta_{01}$ and $\Delta_{21}$. All other quantities are point identified under Assumptions WS-ER. Hence, these equations can be treated as a system of two equations in two unknowns. Letting

$$\begin{aligned} \boldsymbol{A} &= \begin{pmatrix} \omega_{01\cdot 1} & \omega_{21\cdot 1} \\ \omega_{01\cdot 2} & \omega_{21\cdot 2} \end{pmatrix} \\ \boldsymbol{b} &= \begin{pmatrix} q_{11\cdot 11} - q_{11\cdot 01} - p_{01\cdot 01}\omega_{01\cdot 1} - p_{21\cdot 01}\omega_{21\cdot 1} \\ q_{11\cdot 12} - q_{11\cdot 02} - p_{01\cdot 02}\omega_{01\cdot 2} - p_{21\cdot 02}\omega_{21\cdot 2} \end{pmatrix} \end{aligned}$$

and $\boldsymbol{\Delta} = (\Delta_{01}, \Delta_{21})^T$, the system of equations can be written in matrix notation as $\boldsymbol{A}\boldsymbol{\Delta} = \boldsymbol{b}$. The identifiability results for $\Delta_{01}$ and $\Delta_{21}$ with data from two populations are given in the following theorem.

**Theorem 2** *Under Assumptions WS-ER, if the matrix $\boldsymbol{A}$ is invertible, then the causal estimands $\Delta_{01}$ and $\Delta_{21}$ are point identified, with identifying equations given by*

$$\boldsymbol{\Delta} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

*Proof.* The proof of this theorem follows immediately from the invertibility of $\boldsymbol{A}$. ■

We now consider how the point identified $\boldsymbol{\Delta} = (\Delta_{01}, \Delta_{21})^T$ can be estimated, and how inference for the resulting estimator can be performed.

Considering $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{q})$ and $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{q})$ as functions of $\boldsymbol{q}$ suggests the natural estimators $\widehat{\boldsymbol{A}} = \boldsymbol{A}(\hat{\boldsymbol{q}})$ and $\hat{\boldsymbol{b}} = \boldsymbol{b}(\hat{\boldsymbol{q}})$. We can then construct an estimator for $\boldsymbol{\Delta} =$

$(\Delta_{01}, \Delta_{21})^T$ as

$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{b}}.$$

The consistency and asymptotic distribution of $\widehat{\boldsymbol{\Delta}}$ are given in the following theorem.

**Theorem 3** *Under Assumptions 1-6, if $\boldsymbol{A}$ is invertible then*
*a.) $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\Delta}$, and*
*b.) the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$ is given by*

$$\sqrt{n}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \xrightarrow{d} N_2(\mathbf{0}, \nabla(\boldsymbol{A}^{-1}(\boldsymbol{q})\boldsymbol{b}(\boldsymbol{q}))\Sigma_{\boldsymbol{q}}\nabla(\boldsymbol{A}^{-1}(\boldsymbol{q})\boldsymbol{b}(\boldsymbol{q}))^T).$$

*Proof.* a.) By applying Lemmas 1 and 5 from Chapter 1 to the data from each of the two populations, we have that the individual elements of $\widehat{\boldsymbol{A}}$ and $\hat{\boldsymbol{b}}$ converge in probability to their respective population parameters. Thus, by Slutsky's theorem, the invertibility of $\boldsymbol{A}$, and the Continuous Mapping Theorem we have that $\widehat{\boldsymbol{A}}^{-1} \xrightarrow{p} \boldsymbol{A}^{-1}$ and $\hat{\boldsymbol{b}} \xrightarrow{p} \boldsymbol{b}$. Applying Slutsky's theorem again to the full expression yields $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{b}} \xrightarrow{p} \boldsymbol{A}^{-1}\boldsymbol{b}$. The proof is completed by noting that $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\Delta}$.
b.) By Lemma 6 with $k = 2$, the (scaled and centered) combined vector of estimated observed-data quantities for the two populations $\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\Sigma_{\boldsymbol{q}}$. Viewing $\boldsymbol{A}(\boldsymbol{q})$ and $\boldsymbol{b}(\boldsymbol{q})$ as functions of $\boldsymbol{q}$, it follows by a delta method argument that $\sqrt{n}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$ is asymptotically normal with mean $\mathbf{0}$ and variance $\nabla(\boldsymbol{A}^{-1}\boldsymbol{b})\Sigma_{\boldsymbol{q}}\nabla(\boldsymbol{A}^{-1}\boldsymbol{b})^T$.

## *2.6 Identifiability with Data from $k > 2$ Populations under the Exclusion Restriction Assumption*

In this section, we generalize the results obtained in the previous section for a two-site RCT to the case of a multi-site trial with $K > 2$ sites.

Under Assumptions WS-ER, we have the following set of $K$ identifying equations

$$\begin{aligned}
q_{11\cdot 11} - q_{11\cdot 01} - p_{01\cdot 01}\omega_{01\cdot 1} - p_{21\cdot 01}\omega_{21\cdot 1} &= \Delta_{01}\omega_{01\cdot 1} + \Delta_{21}\omega_{21\cdot 1} \\
q_{11\cdot 12} - q_{11\cdot 02} - p_{01\cdot 02}\omega_{01\cdot 2} - p_{21\cdot 02}\omega_{21\cdot 2} &= \Delta_{01}\omega_{01\cdot 2} + \Delta_{21}\omega_{21\cdot 2} \\
&\vdots \quad \vdots \quad \vdots \\
q_{11\cdot 1k} - q_{11\cdot 0k} - p_{01\cdot 0k}\omega_{01\cdot k} - p_{21\cdot 0k}\omega_{21\cdot k} &= \Delta_{01}\omega_{01\cdot k} + \Delta_{21}\omega_{21\cdot k}.
\end{aligned}$$

Using conventions similar to the previous section, this system of $K$ equations can be written in matrix form as

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\Delta}$$

Replacing the point identified quantities in this matrix equation with their corresponding estimators yields the matrix equation $\hat{\boldsymbol{b}} = \widehat{\boldsymbol{A}}\boldsymbol{\Delta}$. With data from $K > 2$ sites, this matrix equation can be viewed as a system of $K$ equations in two unknowns. As this system is overdetermined, we do not attempt to find an exact solution. Instead, we consider estimating $\boldsymbol{\Delta}$ with the least-squares estimator

$$\hat{\boldsymbol{\Delta}}_{LS} = (\widehat{\boldsymbol{A}}^T\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^T\hat{\boldsymbol{b}}$$

which yields an approximate solution to the system of equations. The large sample behavior of this estimator is given in the following theorem.

**Theorem 4** *Under Assumptions WS-ER, if the matrix $\boldsymbol{A}$ is full rank, then*
*a.) the estimator $\hat{\boldsymbol{\Delta}}_{LS}$ is consistent for $\boldsymbol{\Delta}$*

$$\hat{\boldsymbol{\Delta}}_{LS} \xrightarrow{p} \boldsymbol{\Delta}$$

*and*
*b.) $\sqrt{n}(\hat{\boldsymbol{\Delta}}_{LS} - \boldsymbol{\Delta})$ is asymptotically normal with distribution*

$$\sqrt{n}(\hat{\boldsymbol{\Delta}}_{LS} - \boldsymbol{\Delta}) \xrightarrow{d} N_2\left(\mathbf{0}, \nabla((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b})\Sigma_{\boldsymbol{q}}\nabla((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b})^T\right)$$

*Proof.*

*Consistency.*
By applying Lemmas 1 and 5 to the data from each site, it is clear that the individual elements of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{b}}$ all converge in probability to their respective population parameter. By Slutsky's theorem, the assumption about the rank of $\boldsymbol{A}$, and the Continuous Mapping Theorem, it follows that $\hat{\boldsymbol{A}} \xrightarrow{p} \boldsymbol{A}, (\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}})^{-1} \xrightarrow{p} (\boldsymbol{A}^T\boldsymbol{A})^{-1}$, and $\hat{\boldsymbol{b}} \xrightarrow{p} \boldsymbol{b}$. Thus, by applying Slutsky's theorem to the full expression for $\hat{\boldsymbol{\Delta}}_{LS}$ we obtain $\hat{\boldsymbol{\Delta}}_{LS} \xrightarrow{p} (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$. Finally, noting that $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\Delta}$ completes the proof.
*Asymptotic Normality.*
By Lemma 6, $\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q})$ is asymptotically normal with asymptotic variance $\Sigma_{\boldsymbol{q}}$. Viewing $\boldsymbol{A}$ and $\boldsymbol{b}$ as functions of $\boldsymbol{q}$, it follows by a delta method argument that $\sqrt{n}(\hat{\boldsymbol{\Delta}}_{LS} - \boldsymbol{\Delta})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\nabla((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b})\Sigma_{\boldsymbol{q}}\nabla((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b})^T$. ■

The consistency result from Theorem 4 establishes $\hat{\boldsymbol{\Delta}}_{LS}$ as a valid estimator for $\boldsymbol{\Delta}$, while the asymptotic normality result indicates how inference for this estimator can be performed. Given an estimate $\hat{\boldsymbol{q}}$ of the combined vector of within-site observed-data distributions, we can construct an estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\Delta}}_{LS}$ as $\widehat{Var}(\hat{\boldsymbol{\Delta}}) = \nabla((\widehat{\boldsymbol{A}}^T\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^T\hat{\boldsymbol{b}})\widehat{\Sigma}_{\boldsymbol{q}}\nabla((\widehat{\boldsymbol{A}}^T\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^T\hat{\boldsymbol{b}})^T$. Letting $\hat{\sigma}^2_{LS,01}$ and $\hat{\sigma}^2_{LS,21}$ stand for the corresponding estimators of the asymptotic variances associated with $\Delta_{01}$ and $\Delta_{21}$, respectively, the interval $[\hat{\Delta}_{LS,d1} - z_{1-\alpha/2} \times \hat{\sigma}^2_{LS,d1}, \hat{\Delta}_{LS,d1} + z_{1-\alpha/2} \times \hat{\sigma}^2_{LS,d1}]$ will be an asymptotically valid $100 \times (1-\alpha)$ confidence interval for $\Delta_{d1}, d \in \{0, 2\}$. Similar to the asymptotic distributions of the lower and upper bounds given in Chapter 1, then, inference for $\hat{\boldsymbol{\Delta}}_{LS}$ based on large sample approximations involves evaluating conceptually simple but notationally cumbersome partial derivatives. For simplicity, we avoid dealing with them here and instead focus on the non-parameteric bootstrap as our basis for producing confidence intervals for $\Delta_{01}$ and $\Delta_{21}$. Given some large number $B$ of bootstrapped estimates for $\Delta_{01}$ and $\Delta_{21}$ based on resampling observations with replacement, we can construct a $100 \times (1-\alpha)\%$ confidence interval for $\Delta_{d1}, d \in \{0, 2\}$ as $(\Delta^L_{d1}, \Delta^U_{d1}) = (\Delta^*_{d1,\alpha/2}, \Delta^*_{d1,1-\alpha/2})$, where $\Delta^*_{d1,q}$ is the $100 \times q\%$ quantile of the empirical distribution of the $B$ bootstrapped estimates for $\Delta_{d1}$.

## *2.7 Simulations for $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\Delta}}_{LS}$*

In this section, we present simulation results for the estimators $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\Delta}}_{LS}$. We will focus our simulations on estimation and inference for the causal estimand $\Delta_{01}$.

Estimation and inference for $\Delta_{21}$ is conceptually similar, and simulations conducted to assess the behavior of estimators for $\Delta_{21}$ yielded results highly similar to those we present here. Similar to the simulations presented in Chapter 1, we will use these simulations to examine the influence of the parameters $\omega_{01 \cdot k}$ on the performance of the estimators for $\Delta_{01}$.

We will begin by presenting simulation results related to the consistency of the estimators $\hat{\Delta}_{01}$ and $\hat{\Delta}_{LS,01}$. We will then present simulation results assessing the performance of confidence intervals for these estimators under a range of data-generating mechanisms. To assess the consistency of $\hat{\Delta}_{01}$ and $\hat{\Delta}_{LS,01}$, we generated datasets under three different data-generating scenarios. In the first scenario, the proportion of the population at each site belonging to the $(D(0) = 0, D(1) = 1)$ principal stratum was set to be relatively low (less than or equal to 25% of the overall population in each site). In the second scenario, this proportion was set to a medium level (between 50-60% of the overall population in each site). In the third scenario, this proportion was set to a high level (between 75-85% of the overall population in each site). In each scenario, we generated datasets with sample sizes ranging from $n = 300$ to $n = 300,000$ in increments of 300. At each sample size, we simulated a three-site RCT with control group heterogeneity; in each simulation, there were $n/3$ subjects per site. The underlying causal models for each site were chosen in such a way that the exclusion restriction and comparable strata effects assumptions hold. In addition, confounder-style site effects were included by making the causal outcome parameter vectors for Site 2 and Site 3 equal to the causal outcome parameter vector for Site 1 plus an constant. Table 2.1 shows the full causal model for each site under the three scenarios for the underlying data-generating mechanism. The underlying causal model parameters correspond to a causal effect of $\Delta_{01} = .2$ across all sites.

For each combination of data-generating mechanism and sample size, we generated full sets of counterfactuals for the $n/3$ individuals in each site based on the site-specific causal model parameter values. We then generated the observed data for each site by randomly assigning half of the individuals in each site to the $Z = 0$ arm and half of the individuals to the $Z = 1$ arm. We obtained estimates for $\hat{\Delta}_{LS,01}$ based on the observed data from all three sites. Additionally, we obtained estimates for $\hat{\Delta}_{01}$ based on the observed data from the first two sites only. Figure 2.1 shows a trace plot of $\hat{\Delta}_{LS,01}$ as a function of increasing sample size, while Figure 2.2 shows a trace plot of

| Scenario | Site | $\omega_{00}$ | $\omega_{01}$ | $\omega_{11}$ | $\omega_{21}$ | $p_{00\cdot}$ | $p_{01\cdot 0}$ | $p_{01\cdot 1}$ | $p_{11\cdot}$ | $p_{21\cdot 0}$ | $p_{21\cdot 1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Small $\omega_{01}$ | Site 1 | .05 | .19 | .05 | .71 | .1 | .1 | .3 | .2 | .15 | .2 |
| | Site 2 | 0 | .15 | .05 | .8 | .15 | .15 | .35 | .25 | .2 | .25 |
| | Site 3 | .03 | .25 | .05 | .67 | .18 | .18 | .38 | .28 | .23 | .28 |
| Medium $\omega_{01}$ | Site 1 | .05 | .5 | .05 | .4 | .1 | .1 | .3 | .2 | .15 | .2 |
| | Site 2 | 0 | .6 | .05 | .35 | .15 | .15 | .35 | .25 | .2 | .25 |
| | Site 3 | .03 | .55 | .05 | .37 | .18 | .18 | .38 | .28 | .23 | .28 |
| Large $\omega_{01}$ | Site 1 | .05 | .85 | .05 | .05 | .1 | .1 | .3 | .2 | .15 | .2 |
| | Site 2 | 0 | .75 | .05 | .2 | .15 | .15 | .35 | .25 | .2 | .25 |
| | Site 3 | .03 | .8 | .05 | .12 | .18 | .18 | .38 | .28 | .23 | .28 |

Table 2.1: Underlying causal model parameter values for the three sites under the three different data generating scenarios.

$\hat{\Delta}_{01}$ as a function of increasing sample size.

The trace plots in Figures 2.1 and 2.2 demonstrate how both estimators converge to the true underlying value of $\Delta_{01}$ under all three data-generating scenarios we have considered. However, it is also obvious from these plots that the speed of convergence does depend on the underlying values of $\omega_{01\cdot k}$. When $\omega_{01\cdot k}$ is relatively large in all three sites, both estimators converge rapidly to the true value. On the other hand, when $\omega_{01\cdot k}$ is relatively small in all three sites, the estimators continue to exhibit a noticeable amount of variance around the true value of $\Delta_{01}$ even at sample sizes as high as $n = 300,000$. This influence of the parameters $\omega_{01\cdot k}$ on the performance of the estimators mirrors the simulation results given in Chapter 1 in the single-site case.

The simulation results for the two-site estimator $\hat{\Delta}_{01}$ highlight an unfortunate characteristic that the estimators $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\Delta}}_{LS}$ share with method-of-moments estimators, namely that the estimators can sometimes yield values that lie outside of the parameter space. By inspecting the top panel of Figure 2.2, we can see that for small sample sizes the values of $\hat{\Delta}_{01}$ occasionally exceed 1, the maximum value for $\Delta_{01}$. However, this appears to mainly be an issue when small sample sizes are combined with relatively small values for $\omega_{01\cdot k}$. When the values of $\omega_{01\cdot k}$ are in the medium to large range, the estimators exhibit much better behavior.

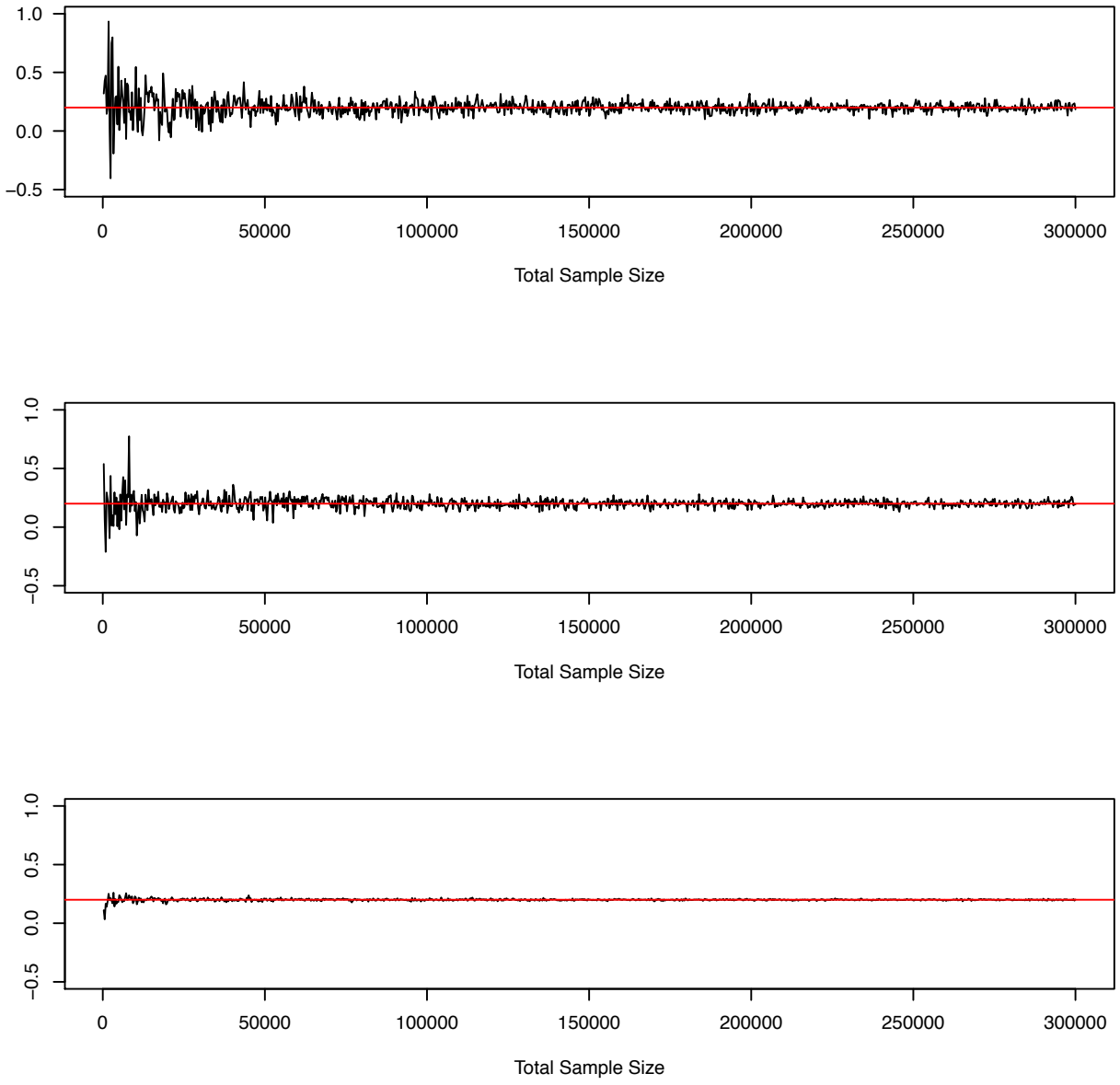



Figure 2.1: Trace plots showing the convergence of the estimator $\hat{\Delta}_{LS,01}$ to the true value of the general multi-site estimator $\Delta_{01}$ (indicated by the red solid line) as the sample size increases. The top panel shows results from the scenario where $\omega_{01}$, the proportion of individuals belonging to the $D(0) = 0, D(1) = 1)$ principal stratum, was set to a relatively low value in all three sites. The middle value shows results from the scenario where $\omega_{01}$ was set to a value around 0.5 in all three sites. The bottom panel shows results from the scenario where $\omega_{01}$ was set to a relatively high value in all three sites.

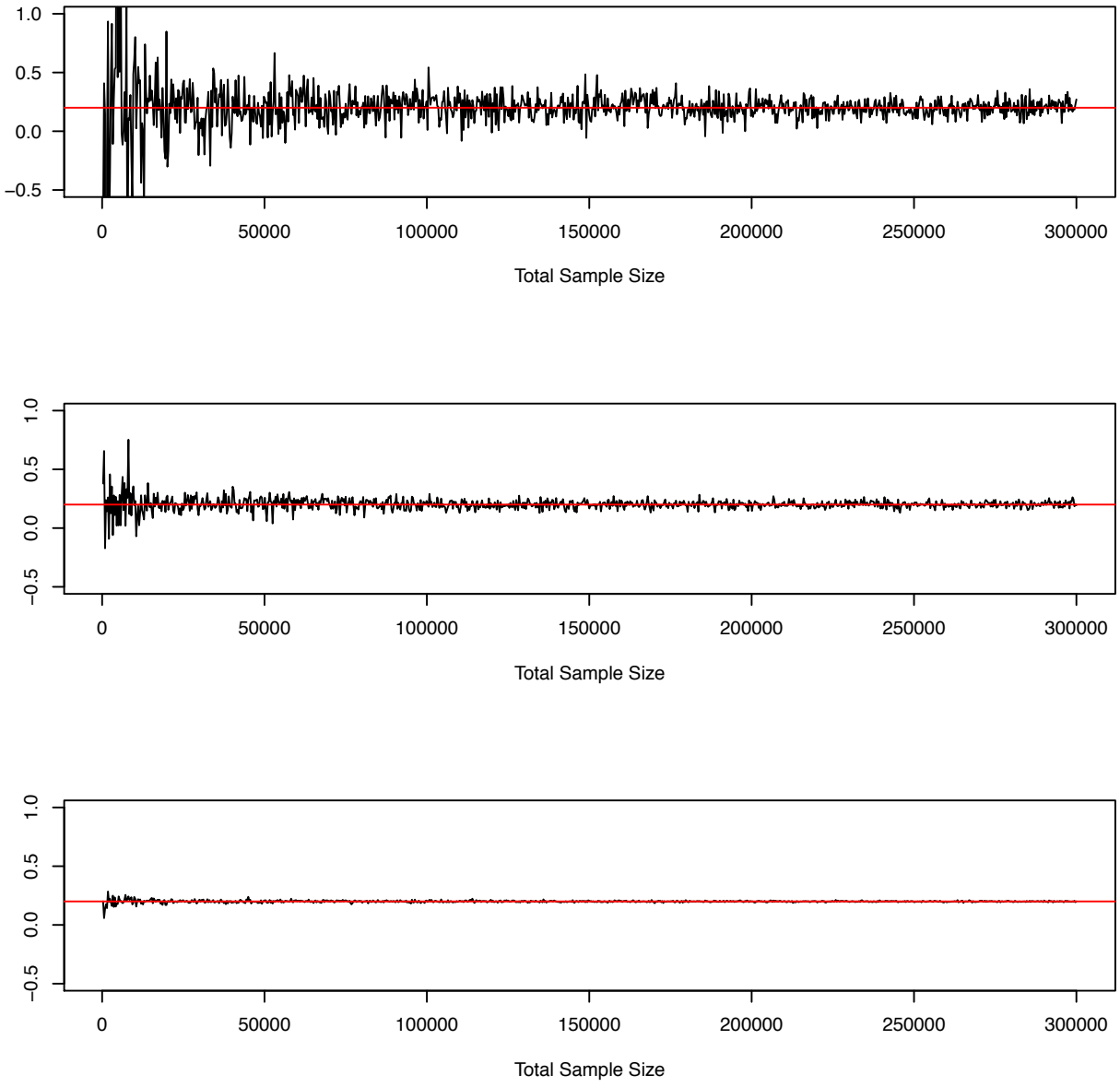



Figure 2.2: Trace plots showing the convergence of the two-site specific estimator $\hat{\Delta}_{01}$ to the true value of $\Delta_{01}$ (indicated by the red solid line) as the sample size increases. The top panel shows results from the scenario where $\omega_{01}$, the proportion of individuals belonging to the $D(0) = 0, D(1) = 1)$ principal stratum, was set to a relatively low value in all three sites. The middle value shows results from the scenario where $\omega_{01}$ was set to a value around 0.5 in all three sites. The bottom panel shows results from the scenario where $\omega_{01}$ was set to a relatively high value in all three sites.

| Total Sample Size | Magnitude of $\omega_{01}$ | Bias | 95% CI Coverage |
|---|---|---|---|
| 300 | Small | -0.07 | 0.99 |
| 300 | Medium | -0.03 | 0.99 |
| 300 | Large | -0.003 | 0.98 |
| 3000 | Small | -0.02 | 0.98 |
| 3000 | Medium | 0.001 | 0.98 |
| 3000 | Large | -0.0002 | 0.95 |
| 30000 | Small | -0.003 | 0.96 |
| 30000 | Medium | 0.001 | 0.95 |
| 30000 | Large | 0.0003 | 0.94 |

Table 2.2: Estimated coverage of 95% confidence intervals for $\Delta_{01}$ with data from multiple populations under Assumptions WS-ER.

We now perform simulations to examine the bias and coverage probabilities for the bootstrap-based confidence intervals for $\widehat{\Delta}_{01,LS}$. We use the same parameter values as in the consistency simulations. Using these parameter values, we generate data sets at sample sizes of $n = 300$, $n = 3000$, and $n = 30000$. In each case, the sample size per site is set to $n/3$. At each sample size and for each data-generating scenario, we generate 1000 datasets. For each dataset, we calculate a point estimate and a 95% bootstrap-based confidence interval for $\widehat{\Delta}_{01,LS}$ based on 1000 bootstrap replications. In Table 2.2, we report estimates of the bias and 95% confidence interval coverage rates at each sample size and data-generating scenario.

When the proportion of the population belonging to the $(D(0) = 0, D(1) = 1)$ principal stratum is high ( 70% or greater), the finite sample bias is negligible even at the lowest sample size setting and continues to decrease as the sample size in each site increases. On the other hand, there is considerable bias when the proportion of the population belonging to the $(D(0) = 0, D(1) = 1)$ is low (bias of around 0.07 compared to a causal effect of 0.20). Thus, even when we have point identification of the causal estimand $\Delta_{01}$, the associated parameter $\omega_{01}$ exerts a noticeable influence on how well estimators of $\Delta_{01}$ will perform in finite samples. Coverage probabilities show a tendency to be conservative at small sample sizes regardless of the proportion of the population belonging to the $(D(0) = 0, D(1) = 1)$ stratum.

## *2.8 Re-analysis of the ESDM RCT data*

Unfortunately, data from a multi-site RCT subject to control group heterogeneity from a TAU control group were not available for analysis. However, our approach to obtaining identifiability of $\Delta_{01}$ and $\Delta_{21}$ is just as valid when conditioning on an appropriately chosen categorical covariate $S$ rather than a variable indicating which site an observation belongs to. Thus, the data do not necessarily have to come from a multi-site RCT in order for our method to apply, so long as Assumptions WS-ER are satisfied for some baseline covariate $S$. We can therefore re-analyze the single-site ESDM RCT data analyzed in Chapter 1 using the techniques developed in this chapter by finding an appropriate covariate and treating the sample strata defined by that covariate as if they came from their own separate sub-populations.

As in the analysis from Chapter 1, we define the analysis variables as follows. The outcome of interest was an increase in the Mullen Composite score, a measure of cognitive development, of at least 15 points at one year post-randomization compared to baseline, so that the outcome $Y$ is an binary variable taking the value of 1 if the Mullen Composite score was 15 points greater at 1 year post-randomization compared to baseline and 0 otherwise. Individuals assigned to the ESDM intervention group ($Z = 1$) were categorized as having received the intervention ($D = 1$) if they received an average of 10 hours or more per week of the ESDM intervention during the intervention period. Individuals assigned to the ESDM intervention group who received fewer than 10 hours per week on average of ESDM intervention were categorized as having not received intervention ($D = 0$). For individuals assigned to the community-based care control group ($Z = 0$), the level of individual treatment was based on the number of hours of individual intensive applied behavioral analysis (ABA) therapy. Children who received greater than 10 hours per week on average of individualized treatment were categorized as having received individual-based treatment ($D = 1$). Children who received fewer than 10 hours per week on average of individualized treatment and greater than 10 hours per week of group treatment were categorized as having received group-based treatment ($D = 2$). Children who received fewer than 10 hours per week of individualized treatment and fewer than 10 hours per week of group treatment were categorized as having not received treatment ($D = 0$).

In addition to these variables from the original analysis, we will also introduce

as our conditioning covariate $S$ an indicator for whether the child was less than 24 months of age at baseline, $S = \mathbb{1}(\text{Age}_{\text{Baseline}} \leq 24 \text{ Months})$. The choice of this variable is motivated by the likely presence of a relationship between the age of a child and the type of community-based treatment that the child's caretaker may choose for him or her. Thus, we would expect the distribution of the principal strata to differ between the $S = 0$ and $S = 1$ groups. The estimated distribution of the principal strata for children greater than 24 months of age at baseline $S = 0$ was estimated as $\hat{\boldsymbol{\omega}}_0 = (0, 0.5, 0.17, 0.33)$. The estimated distribution of the principal strata for children less than 24 months of age at baseline $S = 1$ was estimated as $\hat{\boldsymbol{\omega}}_1 = (0, 0.73, 0.13, 0.13)$. Thus, the distribution of the principal strata does indeed appear to differ between these two subgroups of the overall sample. Children younger than 24 months of age at baseline are more likely to not receive any community-based treatment if assigned to the control group compared to children older than 24 months of age at baseline. Both groups have similar proportions estimated to belong to the $\omega_{11}$ strata, while the older group is estimated to have more children who would primarily receive some form of group-based care if assigned to the treatment group. These differences seem plausible, as older children are more likely to be enrolled in educational programs that may feature some sort of group-based treatment (e.g. special needs classes). We note that the proportion of children who would receive no care under either treatment arm is estimated to be 0 in both groups. This does not affect our analysis, as the positivity assumption on the principal strata parameters is only required for $\omega_{01}$ and $\omega_{21}$.

The observed-data distributions for the two groups $\hat{\boldsymbol{q}}_0$ and $\hat{\boldsymbol{q}}_1$ both pass the model checks for the exclusion restriction assumption given in Chapter 1. Hence, we will proceed under Assumptions WS-ER and compute a single point estimate and bootstrap-based confidence interval for $\Delta_{01}$. The point estimate for $\Delta_{01}$ based on this two-group analysis is $\hat{\Delta}_{01} = 0.47$, with a 95% bootstrap confidence interval of $(-0.84, 1.87)$, based on 10,000 bootstrap resamplings. Similar to the analysis in Chapter 1, this analysis highlights both the potential of our approach and its limitations. In this case, the small sample sizes and the presence of only two levels in the conditioning covariate lead to high variability in the estimate of $\Delta_{01}$. This results in the bootstrap confidence interval extending into values greater than 1, which are not possible for $\Delta_{01}$. The conclusion is that we cannot have much confidence in this particular esti-

mate of $\Delta_{01}$ based on this sample. Still, this analysis serves as a useful illustration of how our method can be applied in practice. As demonstrated in the simulations in Section 2.7, we would expect to obtain a less variable estimator and hence better bootstrap-based confidence intervals with larger samples.

## *2.9 Identifiability with Data from Two Populations without the Exclusion Restriction*

We now consider what identifiability results can be obtained for $\Delta_{01}$ and $\Delta_{21}$ with data from multiple populations after relaxing the exclusion restriction assumption. As in Section 2.5, we will begin with the specific case of a two-site RCT before considering the more general case of multi-site RCT with $K \geq 2$ sites.

In Chapter 1, we demonstrated the non-identifiability of $\Delta_{01}$ and $\Delta_{21}$ in a single-site trial under the exclusion restriction assumption with a numeric counterexample. We now demonstrate the non-identifiability of $\Delta_{01}$ and $\Delta_{21}$ in the case of a single-site RCT without the exclusion restriction. That is, we wish to show that there exist causal model parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ with associated observed-data distributions $\boldsymbol{q}$ and $\boldsymbol{q}^{'}$ such that $\Delta_{01} \neq \Delta_{01}^{'}$ and $\Delta_{21} \neq \Delta_{21}^{'}$ but $\boldsymbol{q} = \boldsymbol{q}^{'}$. The equation $\boldsymbol{q} = \boldsymbol{q}^{'}$ implies the following set of equations relating $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$:

$$
\begin{aligned}
0 &= (1 - p_{00\cdot0})\omega_{00} + (1 - p_{01\cdot0})\omega_{01} - ((1 - p^{'}_{00\cdot0})\omega^{'}_{00} + (1 - p^{'}_{01\cdot0})\omega^{'}_{01}),\\
0 &= p_{00\cdot0}\omega_{00} + p_{01\cdot0}\omega_{01} - (p^{'}_{00\cdot0}\omega^{'}_{00} + p^{'}_{01\cdot0}\omega^{'}_{01}),\\
0 &= (1 - p_1 1\cdot 0)\omega_{11} - ((1 - p^{'}_1 1\cdot 0)\omega^{'}_{11}),\\
0 &= p_{11\cdot0}\omega_{11} - p^{'}_{11\cdot0}\omega^{'}_{11},\\
0 &= (1 - p_{21\cdot0})\omega_{21} - (1 - p^{'}_{21\cdot0})\omega^{'}_{21},\\
0 &= p_{21\cdot0}\omega_{21} - p^{'}_{21\cdot0}\omega^{'}_{21},\\
0 &= (1 - p_{00\cdot1})\omega_{00} - (1 - p^{'}_{00\cdot1})\omega^{'}_{00},\\
0 &= p_{00\cdot1}\omega_{00} - p^{'}_{00\cdot1}\omega^{'}_{00},\\
0 &= p_{01\cdot1}\omega_{01} + p_{11\cdot1}\omega_{11} + p_{21\cdot1}\omega_{21} - (p^{'}_{01\cdot1}\omega^{'}_{01} + p^{'}_{11\cdot1}\omega^{'}_{11} + p^{'}_{21\cdot1}\omega^{'}_{21}),\\
0 &= (1 - p_{01\cdot1})\omega_{01} + (1 - p_{11\cdot1})\omega_{11} + (1 - p_{21\cdot1})\omega_{21} - (1 - p^{'}_{01\cdot1})\omega^{'}_{01} + (1 - p^{'}_{11\cdot1})\omega^{'}_{11} + (1 - p^{'}_{21\cdot1})\omega^{'}_{21}.
\end{aligned}
$$

This is a set of eight equations that are nonlinear in the 24 unknowns. If we set $\boldsymbol{\omega}$

| Causal Parameter Vector | $q_{01\cdot0}$ | $q_{11\cdot0}$ | $q_{10\cdot0}$ | $q_{20\cdot0}$ | $q_{21\cdot0}$ | $q_{00\cdot1}$ | $q_{01\cdot1}$ | $q_{11\cdot1}$ |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\theta}$ | 0.14 | 0.008 | 0.092 | 0.19 | 0.01 | 0.095 | 0.005 | 0.24 |
| $\boldsymbol{\theta}^{'}$ | 0.14 | 0.008 | 0.092 | 0.19 | 0.01 | 0.095 | 0.005 | 0.24 |

Table 2.3: Observed-data distributions implied by the underlying causal parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$. The equality of the two observed-data distributions despite the different underlying values for $\Delta_{01}$ and $\Delta_{21}$ demonstrates the identifiability issue for these causal estimands without the exclusion restriction assumption.

and $\boldsymbol{\omega}^{'}$ to some specific values, then this becomes an underdetermined system of eight linear equations in 16 unknowns. Hence, we are free to set $p_{01\cdot1}, p^{'}_{01\cdot1}, p_{01\cdot0}, p^{'}_{01\cdot0}, p_{21\cdot1}, p^{'}_{21\cdot1}, p_{21\cdot0}$ and $p^{'}_{21\cdot0}$ so that $\Delta_{01} \neq \Delta^{'}_{01}$ and $\Delta_{21} \neq \Delta^{'}_{21}$. We are then left with a system of eight equations in eight unknowns. Solving this reduced system will then lead to causal parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{'}$ with different values for $\Delta_{01}$ and $\Delta_{21}$ but identical observed-data distributions. As a specific numerical example, we can consider the two possible causal outcome parameters $\boldsymbol{\theta} = (0.2, 0.2, 0.08, 0.05, 0.05, 0.3, 0.2, 0.2, 0.1, 0.6, 0.1, 0.2)$ and $\boldsymbol{\theta}^{'} = (0.8, 0.1, 0.08, 0.05, 0.05, 0.3, 0.4, 0.1, 0.1, 0.6, 0.1, 0.2)$. The observed-data distributions implied by these two causal outcome parameters are identical (see Table 2.3), but $\Delta_{01} \neq \Delta^{'}_{01}$ and $\Delta_{21} \neq \Delta^{'}_{21}$.

The causal estimands $\Delta_{01}$ and $\Delta_{21}$ are thus unidentified based on data from only a single site without the exclusion restriction. Of course, these causal estimands are also unidentified based on data from a single-site RCT with the exclusion restriction, but become point identified under certain additional assumptions when data from other sites are available, so long as the exclusion restriction assumption can be made. This is accomplished by exploiting the assumption of common treatment effects across sites along with the exclusion restriction to construct multiple equations in which the only unidentified quantities are the causal estimands $\Delta_{01}$ and $\Delta_{21}$. We now examine why a similar approach cannot work for a multi-site RCT without the exclusion restriction. From Lemma 2 from Chapter 1, we know that under Assumptions WS-NER the unidentified causal parameters are related to the observed-data quantities within each

site by the following equations:

$$\begin{aligned} q_{11\cdot 1k} &= p_{01\cdot 1k}\omega_{01\cdot k} + p_{11\cdot 1k}\omega_{11\cdot k} + p_{21\cdot 1k}\omega_{21\cdot k}, \\ q_{01\cdot 0k} &= p_{00\cdot 0k}\omega_{00\cdot k} + p_{01\cdot 0k}\omega_{01\cdot k} \end{aligned} \tag{2.2}$$

for $k \in \{1, 2\}$. Subtracting $q_{01\cdot 0k} + p_{21\cdot 0k}\omega_{21\cdot k}$ from the first equation in (2.2) gives

$$\begin{aligned} q_{11\cdot 1k} - q_{01\cdot 0k} - p_{21\cdot 0k}\omega_{21\cdot k} &= p_{01\cdot 1k}\omega_{01\cdot k} + p_{11\cdot 1k} + p_{21\cdot 1k}\omega_{21\cdot k} - p_{01\cdot 0k}\omega_{01\cdot k} - p_{00\cdot 0k}\omega_{00\cdot k} - p_{21\cdot 0k}\omega_{21\cdot k} \\ &= \Delta_{01}\omega_{01\cdot k} + \Delta_{21}\omega_{21\cdot k} + p_{11\cdot 1k}\omega_{11\cdot k} - p_{00\cdot 0k}\omega_{00\cdot k}. \end{aligned}$$

This equation shows how point identification for the causal estimands $\Delta_{01}$ and $\Delta_{21}$ is not obtainable with data from two populations without the exclusion restriction assumption. Although the addition of an extra site has provided an additional equation relating $\Delta_{01}$ and $\Delta_{21}$ to observed data quantities, it comes at the price of adding two additional causal parameters $p_{01\cdot 12}$ and $p_{21\cdot 12}$, neither of which is point identified based on the observed data from their respective sites.

## *2.10 Identifiability with Data from $K \geq 2$ Populations without the Exclusion Restriction*

Having considered the case of a two-site study as a motivating example for the continued identifiability issues that come from relaxing the exclusion restriction assumption, we now generalize these results to the case of a multi-site trial with $K \geq 2$ sites.

First, we note that the lack of identifiability for $\Delta_{01}$ and $\Delta_{21}$ based on data from a two-site trial cannot be remedied by adding data from additional sites. The reason for this is that each additional site we add contributes essentially only a single identifying equation relevant to $\Delta_{01}$ and $\Delta_{21}$ but two additional unidentified parameters, namely the parameters $p_{00\cdot 0k}$ and $p_{11\cdot 1k}$. Hence, the number of unidentified parameters will always be greater than the number of identifying equations no matter the number of sites in the RCT.

We now consider whether the additional sources of data in a multi-site RCT can lead to improved upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ under Assumptions WSNER, when compared to the bounds derived in Chapter 1 in the case of a single site RCT without the exclusion restriction assumption. One straightforward approach

would be to apply the bounds derived in the previous chapter to each of the sites separately, and then choose the lower and upper bound that provide the tightest bounds on the causal estimand of interest. Letting $\widehat{\Delta}^L_{d1\cdot k}$ and $\widehat{\Delta}^U_{d1\cdot k}$ stand for the estimated lower and upper bounds on $\Delta_{d1}, d \in \{0, 2\}$ based on the data from the $k^{th}$ site, we simply set $\widehat{\Delta}^L_{d1} = \max_k \widehat{\Delta}^L_{d1\cdot k}$ and $\widehat{\Delta}^U_{d1} = \min_k \widehat{\Delta}^U_{d1\cdot k}$. These bounds are self-evidently at least as good as the bounds that would be estimated from any single site alone (although in small samples we may be concerned with getting a pathological data set in which the maximum of the lower bounds exceeds the minimum of the upper bounds). Furthermore, although the actual estimated values for the bounds appear to only be based on data from at most two of the sites, we note that the entirety of the data from all $k$ sites is used in determining which sites' lower and upper bounds are used. As noted in Koenker [2005], a similar phenomenon occurs in quantile regression; a quantile regression with $p$ regression parameters will necessarily have an exact fit to $p$ points in the data set. As a result, the actual value of the estimate of the quantile regression parameters for a given quantile is essentially determined by those $p$ points. The fact that the full data set contributes to determining which of the $p$ points are included in the exact fit saves the quantile regression estimators from having an effective sample size of $n = p$. Intuitively, then, we might expect that the estimators for the lower and upper bounds on $\Delta_{01}$ and $\Delta_{21}$ just described would have lower variance compared to the lower and upper bounds based on data from any single site alone, even though the actual value for the estimate of the lower/upper bound is based on data from a single site alone.

While these bounds estimators are simple to implement, they have two main drawbacks, both related to the fact that we are effectively estimating $K$ separate sets of bounds. The first drawback, already mentioned, is that in finite samples we may observe data in which the maximum estimated lower bound across all sites actually exceeds the minimum estimated upper bound across all sites. The second drawback is that this approach fails to consider all constraints imposed on $\Delta_{01}$ and $\Delta_{21}$ by the data from all $K$ sites simultaneously. Thus, while these bounds are potentially an improvement on the bounds based on any single site, they are not necessarily the best bounds we can obtain. By instead formulating the bounds as the solutions to a set of linear programming problems, we address both of these drawbacks.

In the case of the linear programming problems considered in Chapter 1, analytic

expressions for the resulting bounds were straightforward to obtain. This is not the case for the upper and lower bounds based on data from a $K$-site RCT, due to the increased number of constraints and variables. It will therefore be necessary to introduce more formally the concept of a linear program and the techniques for solving them. We will then show how the problem of finding bounds on $\Delta_{01}$ and $\Delta_{21}$ with data from a multi-site RCT subject to the constraints implied by Assumptions WS-NER can be expressed in terms of a canonical linear program. Once that is accomplished, algorithms from the linear programming literature can be applied to obtain the bounds.

Optimization problems can be described generally as the problem of finding the maximum/minimum of a function of one or more variables (known as the objective function) subject to a set of constraints on those variables. When both the objective function and the constraints on the variables are linear in the unknown variables, the resulting optimization problem is called a linear program. More specifically, an optimization problem is known as a linear program if it can be expressed in the following canonical form.

*Canonical form linear programming problem.* (Adapted from Moutesek and Gartner [2005])
Let $\boldsymbol{x}$ be an $n-$element vector of non-negative variables, $A$ be an $m \times n$ matrix of known real-valued coefficients, and finally let $\boldsymbol{c}$ and $\boldsymbol{b}$ be $n-$ and $m-$ element real-valued vectors, respectively. A canonical form linear program is an optimization problem that has the following form:

$$\begin{aligned} &\text{Maximize } \boldsymbol{c}^T\boldsymbol{x} \\ &\text{subject to } A\boldsymbol{x} = \boldsymbol{b} \\ &\boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

Here, $\boldsymbol{c}^T\boldsymbol{x}$ describes the objective function, and $\boldsymbol{A}$ and $\boldsymbol{b}$ together describe the constraints on the (non-negative) unknown variables $\boldsymbol{x}$, typically referred to as the decision variables.

Once a problem is stated in canonical form, there are many algorithms available to

compute a solution. Of course, many problems of interest do not immediately conform to this canonical form. Decision variables may be unrestrained in sign, for instance, or some constraints may be formulated in terms of inequalities rather than equalities. Luckily, there are numerous ways to transform non-canonically stated problems into canonical form (Moutesek and Gartner [2005] and Sallan et al. [2015] provide a wide range of examples).

In many situations, the decision variables represent the only true unknowns, while the coefficients correspond to some known value. In a business context, for instance, the decision variables may correspond to the number of units of different products to produce. The constraints on these variables would then be based on known production costs and budget constraints. In our setting, the "true" underlying linear program will depend completely on the unknown model parameters. However, the identification results from Chapter 1 give us guidance as to which variables must be treated as the unknown decision variables and which can be treated as the known coefficients or constants in the constraints. The decision variables correspond to the causal parameters that are unidentified under Assumptions WS-NER. Any causal parameter or observed-data quantity that is point identified under Assumptions WS-NER can be treated as a known quantity in the linear program.

We now describe how the problem of bounding $\Delta_{01}$ and $\Delta_{21}$ based on data from a $K$-site RCT can be put into the form of a canonical linear program. Given data from $K \geq 2$ sites, we have $K$ equations (one per site) of the form

$$q_{11\cdot 1k} - q_{01\cdot 0k} - p_{21\cdot 0k}\omega_{21\cdot k} = \Delta_{01}\omega_{01\cdot k} + \Delta_{21}\omega_{21\cdot k} + p_{11\cdot 1k}\omega_{11\cdot k} - p_{00\cdot 0k}\omega_{00\cdot k}.$$

The left-hand side of this equation consists of quantities that are all point identified for each site under Assumptions WS-NER. On the right hand side, the quantities $\Delta_{01}$, $\Delta_{21}$, $p_{11\cdot 1k}$ and $p_{00\cdot 0k}$ are unidentified under Assumptions WS-NER. However, the coefficients associated with these unidentified parameters, namely the site-specific distribution of principal strata $\boldsymbol{\omega}_k$, are point identified under these assumptions. In

addition, the unidentified parameters are constrained by the following inequalities:

$$\begin{aligned} -1 \leq \Delta_{01} \leq 1, \\ -\min_k p_{21\cdot 0k} \leq \Delta_{21} \leq 1 - \max_k p_{21\cdot 0k} \\ 0 \leq p_{11\cdot 1k}, p_{00\cdot 0k} \leq 1, \end{aligned}$$

where the tighter bounds on $\Delta_{21}$ follow from the fact that the causal parameters $p_{21\cdot 0k}$ are point identified under Assumptions WS-NER. Importantly, these equations and inequalities are linear in the unidentified parameters. Thus, the problem of finding bounds on $\Delta_{01}$ and $\Delta_{21}$ under Assumptions WS-NER can be viewed as a set of linear programs with $2K + 2$ decision variables $\Delta_{01}, \Delta_{21}, p_{11\cdot 1k}, p_{00\cdot k}, k \in \{1, 2, \ldots, K\}$ and constraints determined by the observed-data distributions $\boldsymbol{q}_k$ and the natural lower and upper bounds on probabilities and differences in probabilities. This formulation of the linear program is not in canonical form due to $\Delta_{01}$ and $\Delta_{21}$ being unrestrained in sign and some of the constraints being in the form of inequalities rather than equalities. However, it can easily be put into canonical form as follows. First, we replace $\Delta_{01}$ and $\Delta_{21}$ with the non-negative decision variables $\Delta_{01}^{'}, \Delta_{01}^{''}, \Delta_{21}^{'}, \Delta_{21}^{''}$, which relate to the original decision variables through the relationships $\Delta_{01} = \Delta_{01}^{'} - \Delta_{01}^{''}$ and $\Delta_{21} = \Delta_{21}^{'} - \Delta_{21}^{''}$. Then, for the decision variables $p_{00\cdot 0k}, p_{11\cdot 1k}, k \in \{1, 2, \ldots, K\}$ and $\Delta_{d1}^{'}, \Delta_{d1}^{''}, d \in \{0, 2\}$ we add corresponding slack variables $p_{00\cdot 0k}^{'}, p_{11\cdot 1k}^{'}, \delta_{d1}^{'}$, and $\delta_{d1}^{''}$. The canonical form of the bounds problem for $\Delta_{01}$ and $\Delta_{21}$ is then given by the following:

Maximize $a(\Delta_{d1}^{'} - \Delta_{d1}^{''})$

subject to

$$
\begin{aligned}
& q_{11\cdot 1k} - q_{01\cdot 0k} - p_{21\cdot 0k}\omega_{21\cdot k} = \Delta_{01}\omega_{01\cdot k} + \Delta_{21}\omega_{21\cdot k} + p_{11\cdot 1k}\omega_{11\cdot k} - p_{00\cdot 0k}\omega_{00\cdot k}, \\
& p_{00\cdot 0k} - p_{00\cdot 0k}^{'} = 1, \\
& p_{11\cdot 1k} - p_{11\cdot 1k}^{'} = 1, \\
& \Delta_{d1}^{'} - \Delta_{d1}^{''} - \delta_{d1}^{'} = 1, \\
& \Delta_{d1}^{''} - \Delta_{d1}^{'} - \delta_{d1}^{''} = 1, \\
& p_{01\cdot 0k}, p_{01\cdot 1k}, p_{21\cdot 1k} \geq 0, \\
& \Delta_{d1}^{'}, \Delta_{d1}^{''}, d_{d1}^{'}, d_{d1}^{''} \geq 0,
\end{aligned}
$$

for $k \in \{1, 2, \ldots, K\}$ and $d \in \{0, 2\}$. Setting $a = 1$ gives the linear program for finding the upper bound on $\Delta_{d1}$, while setting $a = -1$ gives the linear program for finding the lower bound on $\Delta_{d1}$. Having put the linear program for bounding $\Delta_{d1}$ into canonical form, well-known algorithms can now be applied to find solutions. Of course, the values of the constraints in the "true" linear programs determining the true lower and upper bounds on $\Delta_{01}$ and $\Delta_{21}$ based on data from a $K-$site RCT without the exclusion restriction are based on point identified but still unknown parameters. We obtain estimates for these lower and upper bounds on $\Delta_{01}$ and $\Delta_{21}$ by replacing the point identified quantities in the linear program by their corresponding estimators and then solving this modified linear program. For $d \in \{0, 2\}$, we will let $\widehat{\Delta}_{d1}^{L,MS}$ and $\widehat{\Delta}_{d1}^{U,MS}$ represent the resulting lower and upper bound, respectively, on $\Delta_{d1}$, where the superscript $MS$ indicates that these bounds are based on data from a Multi-Site RCT. In all that follows, we will use the `lpSolve` package in `R` to obtain solutions to the specific linear programs we will consider. The `lpSolve` package uses the revised simplex method, a more efficient but mathematically equivalent implementation of Dantzig's simplex method [Nash, 2000]

| Site | $p_{00\cdot 0}$ | $p_{01\cdot 0}$ | $p_{11\cdot 0}$ | $p_{21\cdot 0}$ | $p_{00\cdot 1}$ | $p_{01\cdot 1}$ | $p_{11\cdot 1}$ | $p_{21\cdot 1}$ | $\omega_{00}$ | $\omega_{01}$ | $\omega_{11}$ | $\omega_{21}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | 0.1 | 0.1 | 0.2 | 0.15 | 0.15 | 0.3 | 0.25 | 0.2 | 0.05 | 0.85 | 0.05 | 0.05 |
| Site 2 | 0.15 | 0.15 | 0.25 | 0.2 | 0.2 | 0.35 | 0.3 | 0.25 | 0.04 | 0.75 | 0.05 | 0.16 |
| Site 3 | 0.13 | 0.13 | 0.23 | 0.18 | 0.18 | 0.33 | 0.28 | 0.23 | 0.05 | 0.85 | 0.05 | 0.05 |

Table 2.4: True values for the causal parameter vectors for a three-site randomized trial with multiple versions of control. The values of the causal parameter vectors were set so that Assumptions WS-NER hold.

## *2.11 Simulations for Multi-site Bounds on $\Delta_{01}$ and $\Delta_{21}$ without the Exclusion Restriction*

We now examine through simulations what sort of improvement we can see in the estimators for these bounds. The simulation setup is as follows. We will simulate data from a three site RCT. The causal outcome parameter values for the three sites will be equal to those given in Table 2.4, so that the exclusion restriction assumption no longer holds and therefore $\Delta_{01}$ and $\Delta_{21}$ are no longer point identified. For each scenario, we simulate 1000 data sets. For each data set, we calculate the following: the multi-site bounds on $\Delta_{01}$ based on the data from all three sites and the single-site bounds on $\Delta_{01}$ given in Chapter 1 based on data from each of the individual sites. We will let $\widehat{\Delta}_{01}^{L,MS}$ and $\widehat{\Delta}_{01}^{U,MS}$ represent the lower and upper linear programming-based multi-site bounds on $\Delta_{01}$.

Figures 2.3 - 2.5 show the average bound widths for the multi-site bounds and each of the single-site bounds for the high, medium, and low $\omega_{01}$ scenarios, respectively. From these figures, we can see that having data from multiple sites leads to a considerable tightening of the bounds on $\Delta_{01}$ in all three data generating scenarios. In some cases, we obtain a reduction in average bound width of nearly 50% relative to the tightest single site bounds. Furthermore, the tightening of the bounds occurs even at small sample sizes. The influence of the $\omega_{01}$ parameter on the width of the bounds is also apparent. When $\omega_{01}$ is large in value, the resulting bound width can be as low as 0.10, a reasonably tight bound considering that the exclusion restriction has not been made. However, when $\omega_{01}$ is in the low to medium range, the bounds can become very wide (although we note that even in the low $\omega_{01}$ scenario, the multi-site bounds are still tight enough to determine the direction of the effect).

| $\omega_{01}$ Magnitude | Sample Size per Site | Bias($\widehat{\Delta}_{01}^{L,MS}$) | Bias($\widehat{\Delta}_{01}^{U,MS}$) |
|---|---|---|---|
| Small | 100 | 0.098 | 0.102 |
| | 300 | 0.076 | 0.078 |
| | 500 | 0.071 | 0.067 |
| | 1000 | 0.051 | 0.054 |
| | 3000 | -0.024 | -0.030 |
| Medium | 100 | 0.117 | 0.121 |
| | 300 | 0.095 | 0.097 |
| | 500 | 0.090 | 0.086 |
| | 1000 | 0.070 | 0.073 |
| | 3000 | 0.030 | 0.026 |
| Large | 100 | 0.040 | 0.030 |
| | 300 | 0.022 | 0.020 |
| | 500 | 0.017 | 0.018 |
| | 1000 | 0.015 | 0.013 |
| | 3000 | -0.0001 | -0.0003 |

Table 2.5: Estimated empirical bias of $\widehat{\Delta}_{01}^{L,MS}$ and $\widehat{\Delta}_{01}^{U,MS}$ at different sample sizes and for different values of $\omega_{01}$

.

Of course, tighter estimated bounds are not useful in and of themselves unless the estimated bounds are relatively unbiased for the true bounds. In Table 2.5, we show the empirical bias for $\widehat{\Delta}_{01}^{L,MS}$ and $\widehat{\Delta}_{01}^{U,MS}$ for the three data generating scenarios at sample sizes of $n = 100, 300, 500, 1000$ and 3000. From these estimates, we can see that when $\omega_{01}$ is in the high range, the finite sample empirical bias in the lower and upper bounds for $\Delta_{01}$ is not too severe, and there is a clear trend towards decreasing empirical bias with increasing sample size. When $\omega_{01}$ is in the low to medium range, the finite sample bias is worse. This is especially true at small sample sizes, where the empirical bias can be considerable.

These simulations demonstrate the advantage that can be gained in terms of tightening the width of the bounds by having data available from multiple populations. Once again, we see that the parameter $\omega_{01}$ exerts a considerable influence on both the usefulness of the bounds in terms of their width, as well as the performance of the bounds estimators in finite samples.

The results shown in this section are conditional on the linear programming-based

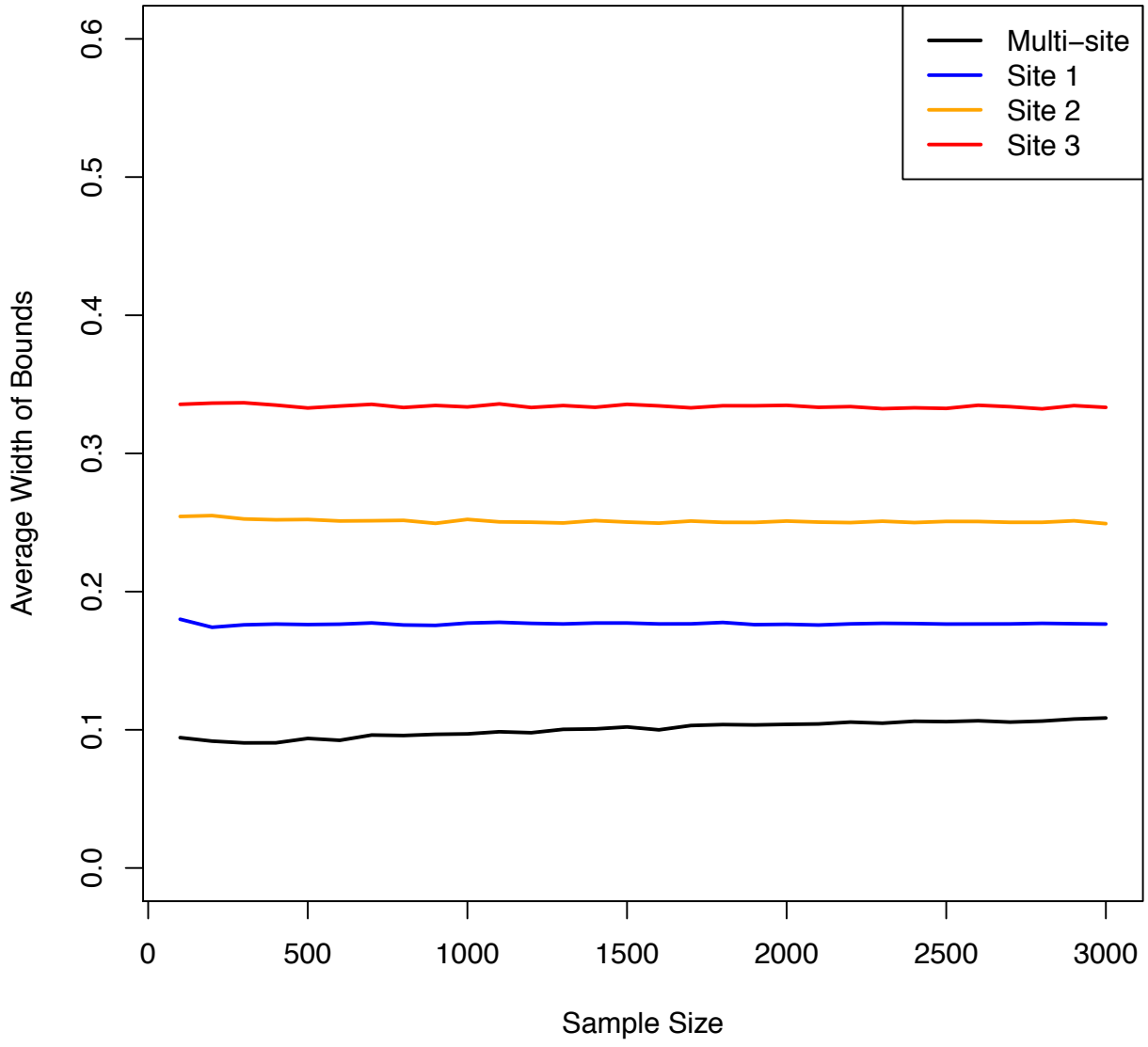



Figure 2.3: Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is relatively large within each site.

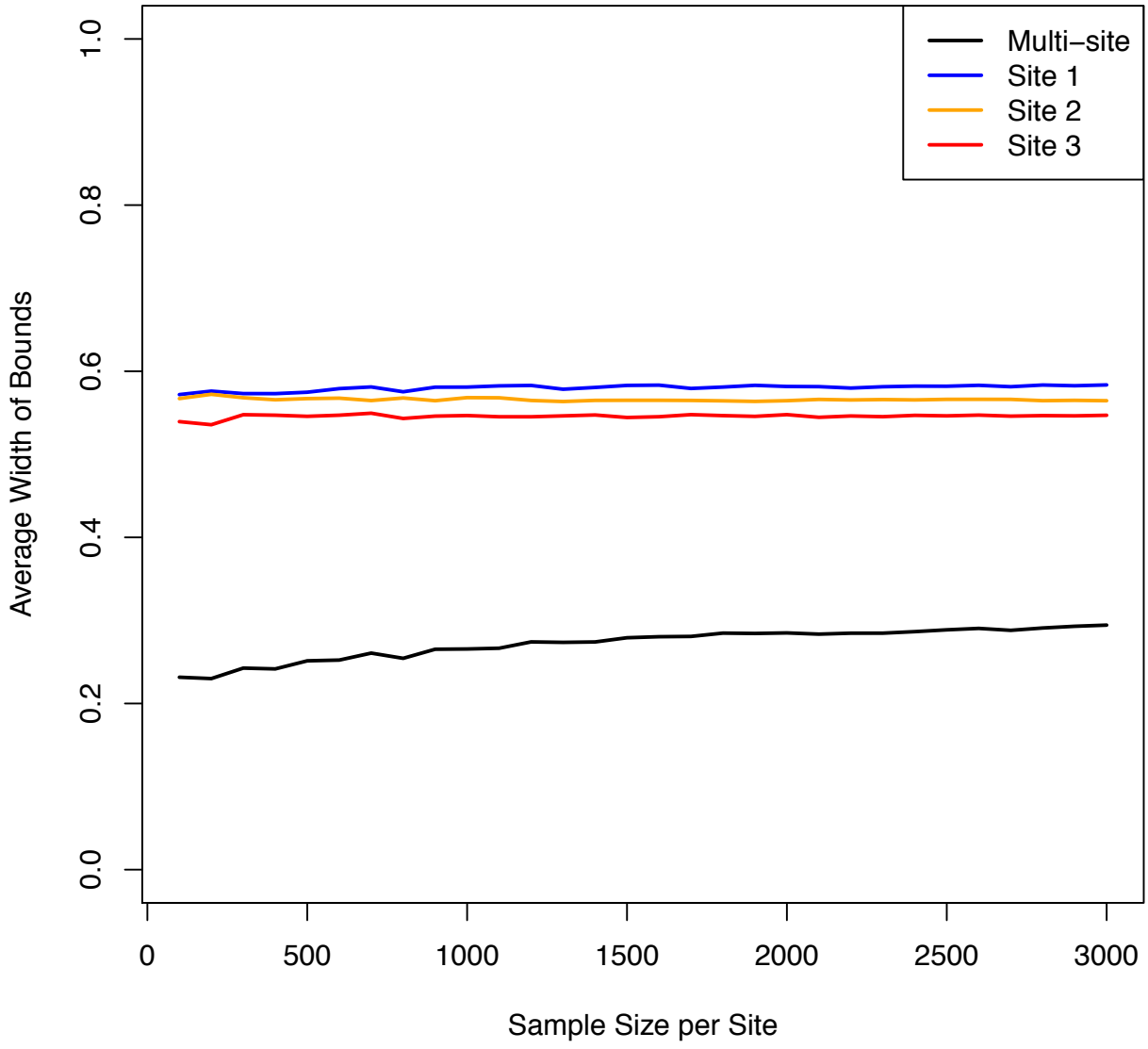



Figure 2.4: Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is in the medium range within each site.

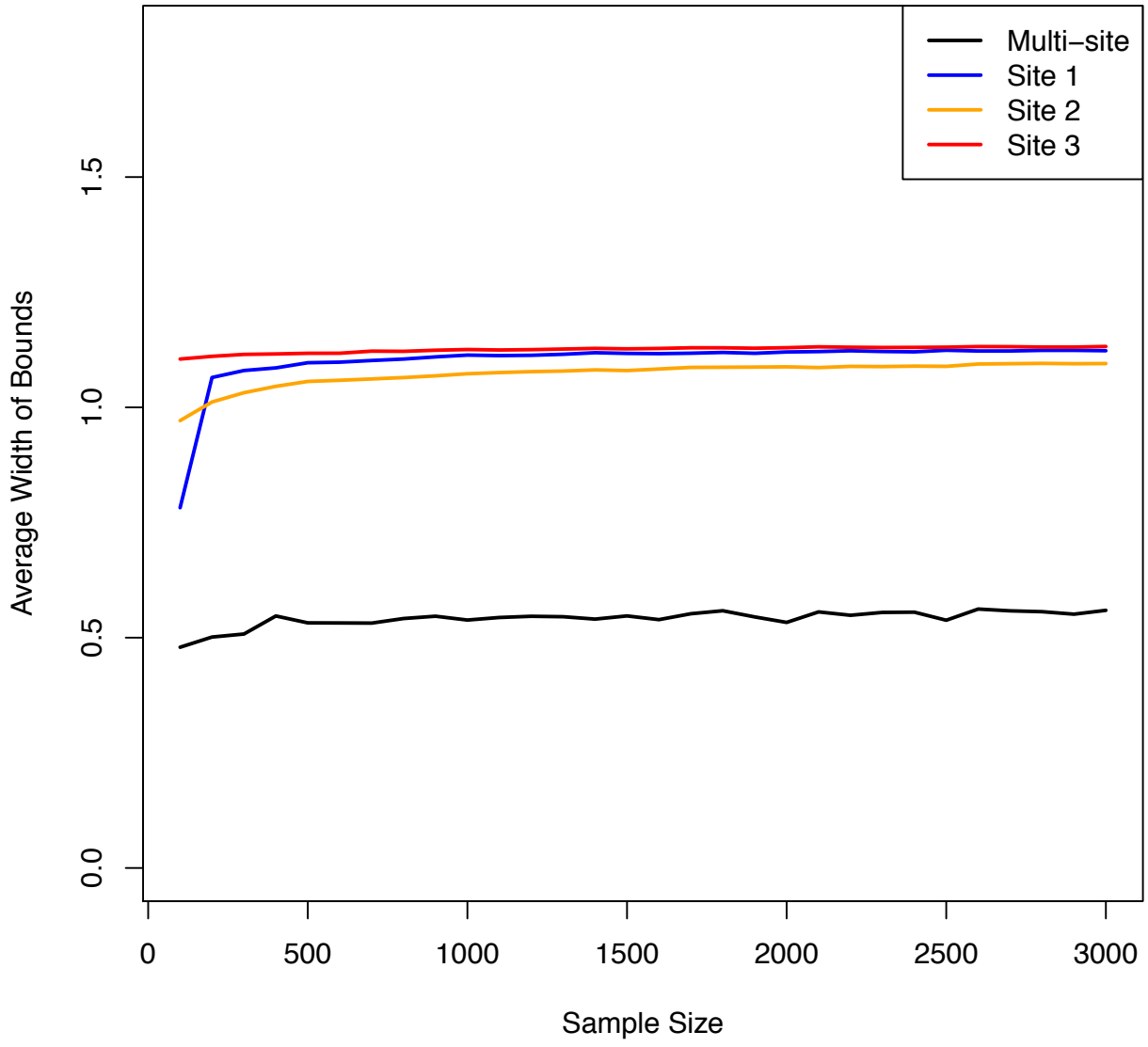



Figure 2.5: Average width of the identified region when the magnitude of the parameters $\omega_{01}$ is relatively small within each site.

approach giving a non-degenerate solution. Unfortunately, there is a tendency for the linear program as currently implemented to produce degenerate solutions, in which the upper and lower bounds are both estimated to be 0. Based on simulations, the probability of a degenerate solution of this type goes to 0 as the sample size increases. The exact reason for these degenerate solutions is not currently known and needs further investigation. If, for a particular sample, the linear programming-based bounds give a degenerate solution, then the bounds discussed at the beginning of Section 2.10 can be used as an alternative.

## *2.12 Comparable Strata Effects Assumption and Treatment Effect Heterogeneity*

The identifiability results obtained in this chapter depend crucially on the Comparable Strata Effects assumption. A natural question is how reasonable the Comparable Strata Effects assumption is likely to be in practice, in particular the assumption of comparability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$ between the two sites. The assumption that the distributions of the principal strata vary across sites seems likely to hold in practice given the natural variability in the availability of community-based options that one would expect to see in different geographic areas. The assumption that $\Delta_{01}$ and $\Delta_{21}$ are comparable between sites, on the other hand, requires more careful consideration. This assumption is closely connected to the question of how to model site effects, a question that has received considerable attention in the literature on the analysis of multi-site RCTs (see, for example, Fleiss [1986], Kraemer and Robinson [2005], Friedman et al. [2015],and Localio et al. [2001]). The crux of the matter is whether site effects should be modeled simply as confounders of the treatment/outcome relationship, or whether possible treatment-by-site variability in the effect of the intervention should be allowed for. In regression terms, the question can be viewed as an issue of whether a main effects model or an interaction model should be used. The advantage of the main effects or simple confounding modeling approach is that the intervention effect can be summarized by a single parameter once adjustment for site has been made. However, the appropriateness of this approach relies on there being true uniformity of the intervention effect across sites. If, on the other hand, intervention effect variability exists between sites, then representing the intervention effect in terms of a single parameter is no longer appropriate. Our

formulation of the Comparable Strata Effects assumption corresponds to the main effects/simple confounding approach to modeling site effects. That is, our approach assumes that there is a single intervention effect across sites for each of the relevant principal strata. Thus, a careful consideration of the potential for site-level variability in the effect of treatment should be made prior to using the approach detailed here. Unfortunately, checking this assumption with the observed data is difficult, as separate point estimates for $\Delta_{01}$ and $\Delta_{21}$ from each site cannot be directly computed. Thus, we cannot simply inspect site-specific estimates of $\Delta_{01}$ and $\Delta_{21}$ to assess how greatly they vary. If data from enough sites are available, one potential data-based approach to assessing the Comparable Strata Effects assumption would be to compute estimates of $\Delta_{01}$ and $\Delta_{21}$ for each pairwise combination of sites. Under the Comparable Strata Effects assumption, each of these point estimates is estimating the same underlying causal estimand. As a result, these point estimates should not vary too greatly. If too much variability in the estimates is observed, then the Comparable Strata Effects assumption may be rejected as unreasonable. As an example, if data from $k = 4$ sites are available, then there are $\binom{4}{2} = 6$ distinct point estimates for $\Delta_{01}$ and $\Delta_{21}$ that can be computed. These estimates could be inspected to see to what extent they agree with each other. A formal test for heterogeneity of treatment effects, however, would likely be difficult to derive and lacking in power, due to the fact that estimates using data from the same site would be correlated.

One point on which the literature seems unified is that the impact of having data from different sites should not be simply ignored in the analysis. By allowing the causal outcome parameters $p_{dy \cdot z}$ to vary across sites, we allow for confounder-type site effects and thus avoid this potential pitfall.

Interestingly, the Comparable Strata Effects assumption, while ruling out treatment effect heterogeneity within the principal strata, does not necessarily rule out treatment effect heterogeneity for the overall $ITT$ effect. The following decomposition of $ITT_k$, the $ITT$ effect within the $k^{th}$ site, makes this clear:

$$\begin{aligned} ITT_k &= P(Y = 1 \mid Z = 1) - P(Y = 1 \mid Z = 0) \\ &= P(Y(1) = 1) - P(Y(0) = 1) \\ &= \sum_{(d_0, d_1) \in \mathcal{D}} \Delta_{d_0 d_1} \omega_{d_0 d_1 \cdot k}. \end{aligned}$$

Under the Comparable Strata Effects assumption, for any two sites $k, k'$, the distributions of principal strata $\omega_{d_0 d_1 \cdot k}$ and $\omega_{d_0 d_1 \cdot k'}$ will differ. Hence, $ITT_k$ will only equal $ITT_{k'}$ in situations where the summands in $ITT_k$ combine in such a way as to equal the total of the summands in $ITT_{k'}$. Otherwise, the $ITT$ effects will differ across sites and thus treatment effect heterogeneity will be present marginally, though not conditionally within principal strata. Thus, we would not be surprised to see heterogeneity between the sites in terms of the estimated $ITT$ effects, even when the Comparable Strata Effects assumption holds. In this way, the Comparable Strata Effects assumption can be viewed as an assumption that the heterogeneity in marginal treatment effects can be attributed to differences in the distribution of community-based care options.

The importance of assuming both common causal effects within sites and different principal strata proportions across sites was recently noted in Jiang et al. [2016], who considered using a multi-site RCT to obtain identifiability of principal causal effects for a binary-valued surrogate endpoint. We note that the development of our framework and the assumptions necessary for it to work occurred independently of their work. It was encouraging to see that similar results regarding the necessity and implications of our assumptions were discovered in other contexts as well. In their work, the authors explore departures from the Comparable Strata Effects assumption as part of a sensitivity analysis. Such an approach could be likely be applied to our framework as well.

## *2.13 Discussion*

The point identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$ obtained in Sections 2.5 and 2.6 is based on the assumption that the distribution of community-based treatment options, which leads to the control group heterogeneity, varies across sites. As noted in Theorems 2 and 4, there are additional restrictions as well on the distributions of the principal strata within each site that must be satisfied in order to obtain identifiability. Thus, the model expansion approach we have considered does not produce identifiability over the full parameter space of the expanded model, but rather a subspace of that parameter space. This finding is consistent with the findings of Gustafson [2005a], who referred to this phenomenon as *essential identifiability* to distinguish it from the standard definition of identifiability.

While we have framed the development of this model expansion approach in the context of a multi-site RCT, we noted in our re-analysis of the ESDM data that it could alternatively have been developed in the context of performing subgroup analyses given data from a single population after conditioning on an appropriately chosen covariate. That is, the results we have obtained do not inherently depend on the data arising from a multi-site trial. The data could instead come from a single-site trial, so long as an appropriate baseline covariate $S$ is available such that the assumptions detailed in Section 2.3 are satisfied after conditioning on $S$. We have chosen to motivate this work specifically in the context of a multi-site trial because it is a common choice of study design in which the necessary assumptions for the identifiability results we have obtained in this chapter seem fairly reasonable (e.g. variability of principal strata distributions across sites). The role of covariates in obtaining tighter bounds on partially identified principal strata effects has been explored elsewhere in the literature. Long and Hudgens [2013] considered whether a baseline categorical variable can sharpen the bounds on principal strata effects when the post-randomization variable is binary. They concluded that incorporating covariate information into the bounds can lead to significant reduction in the width of the identified region, a finding that aligns with the findings of this chapter.

The results in this chapter demonstrate how vital the exclusion restriction assumption is to obtaining identifiability of the causal estimands $\Delta_{01}$ and $\Delta_{21}$. Without the exclusion restriction assumption, these causal estimands remain unidentified even with the addition of data from arbitrarily many populations. Still, even without the exclusion restriction assumption we are able to exploit the additional data from a multi-site study to obtain better bounds on these unidentified estimands. Considering the case where the exclusion restriction fails to hold is especially important in the context of RCTs using a TAU control condition, as the lack of blinding procedures will make the exclusion restriction assumption less likely to hold compared to blinded RCTs. We have shown how model checks for the exclusion restriction assumption can be performed, so that the sample data can at least be checked to see if they are consistent with this assumption. Even if the data do not directly contradict the exclusion restriction assumption, however, investigators may feel uncomfortable making this assumption given the aforementioned lack of blinding procedures. One potential compromise would be to make the exclusion restriction for only one of the

principal strata to which it could be applied. As we have formulated it, the exclusion restriction assumption applies both to the $(D(0) = 0, D(1) = 0)$ stratum and the $(D(0) = 1, D(1) = 1)$ stratum. One area for future research would be to investigate whether similar results are obtainable under a weaker version of this assumption, in which an exclusion restriction assumption is only made about one of the strata. For instance, we might choose to make the exclusion restriction assumption about the $(D(0) = 0, D(1) = 0)$ stratum but allow for effects of intervention assignment in the $(D(0) = 1, D(1) = 1)$ stratum. Such an approach was examined in Hirano et al. [2000] in the case of binary non-compliance in a randomized encouragement design, due to a belief that the exclusion restriction seemed more justifiable for the Never-Taker stratum than the Always-Taker stratum. Similarly, in the framework we have developed here it is arguably more likely that the exclusion restriction holds for those who would never receive treatment under either treatment assignment rather than those who would receive some form of the intervention of interest under either treatment assignment. This is especially true if those in the control group who are categorized as having received the equivalent of the intervention of interest have in fact only received a type of treatment that is highly similar to the intervention of interest.

Unlike the lower and upper bounds on $\Delta_{01}$ and $\Delta_{21}$ derived in Chapter 1, the lower and upper bounds considered in this chapter do not have an easily derived analytic expression, and formal results for the consistency of the associated estimators of these lower and upper bounds have not yet been established. While the simulation results presented in this chapter are suggestive that these lower and upper bounds based on multi-site data can be consistently estimated, further research is needed to confirm this.

Even when enough assumptions hold as to make the causal estimands $\Delta_{01}$ and $\Delta_{21}$ point identified, the ability to obtain good estimates of these quantities, especially at lower sample sizes, depends greatly on their associated principal strata parameter being large enough. If the associated principal strata parameter is too small, then larger samples sizes are needed to obtain convergence and eliminate bias. This makes sense, as these principal strata parameters act as indexes of how much information we would expect there to be about their associated causal estimand in any given sample. Thus, we should not expect to be able to make precise inference about these

causal estimands unless the versions of control associated with them are present in the underlying population at sufficient levels.

Throughout this chapter, we have made a simplyfing assumption of independence between subjects both between and within sites. A natural extension of this approach would be to allow for correlation between individuals in the same site.

# Chapter 3

# CAUSAL INFERENCE WITH PARTIAL COMPLIANCE TO TREATMENT

## *3.1 Introduction*

Causal inference for randomized trials with binary non-compliance is a well-explored topic in the causal inference literature. Inference has often focused on estimation of the Complier Average Causal Effect (CACE), defined as the difference in counterfactual means among individuals whose treatment assignment matches their compliance behavior under both assignment to the control arm and the intervention arm. Angrist et al. [1996] showed how the CACE can be identified by treating the randomized treatment assignment as an instrumental variable. Imbens and Rubin [1997] and Hirano et al. [2000] considered Bayesian approaches to the analysis of RCTs with binary non-compliance. Zhou and Li [2006] and Taylor and Zhou [2009] developed methods for estimation of the CACE in the presence of missing data.

In contrast to the well-explored literature on binary non-compliance, methods for causal inference in the presence of partial compliance have not received as much attention in the literature. Interestingly, one of the earliest papers to explore the role of noncompliance as an explanatory variable did in fact consider the case of partial compliance [Efron and Feldman, 1991]. However, the approach used in Efron and Feldman [1991] relied on assuming a deterministic relationship between the potential compliance behaviors. In this way, the missing compliance behaviors could be directly imputed. More recent papers have focused on estimating partial compliance-based causal effects under less restrictive assumptions. In Jin and Rubin [2008], the authors proposed a Bayesian framework for inference under monotonicity assumptions. Two recent papers [Bartolucci and Grilli, 2011] [Ma et al., 2011] have sought to relax assumptions on the joint distribution of the potential compliance behaviors even further. Both papers used copula-based approaches to model the joint distribution of the potential compliance behaviors. Bartolucci and Grilli [2011] considered the case

of a continuous outcome using a Normal model, with a Plackett copula to model the potential compliance behaviors. Ma et al. [2011] looked at a general approach for outcomes from exponential family distributions, but focused on the case of a binary outcome for simulations. In both cases, inference for the causal parameters of interest was carried out using two-stage procedures. Parameters related to the marginal distributions of the potential compliance behaviors were first estimated, then treated as fixed during estimation of the causal parameters related to the outcome.

Missing from these previous approaches to partial compliance models are the sort of rigorous identification results for causal estimands that are found in the binary non-compliance literature. In this chapter, we cast a critical eye on these previously published methods. We demonstrate that likelihood-based approaches such as those proposed in Ma et al. [2011] and Bartolucci and Grilli [2011] lead to estimators that are either inconsistent for the desired causal estimand even under no model misspecification or almost certain to be biased when used in practice.

### *3.2 Notation and Assumptions for Counterfactuals*

In the following we assume that we have a sample of $n$ individuals taking part in a randomized trial. We let $Y_i$ represent the observed outcome and $Z_i$ represent the randomized treatment assignment for the $i^{th}$ individual. We consider the scenario where individuals in the study may fail to comply with their treatment assignment. We let $D_i$ represent the observed compliance behavior for the $i^{th}$ individual. In contrast to binary noncompliance, where $D_i \in \{0, 1\}$, we consider the case of partial compliance. Under partial compliance, the compliance behavior represents the proportion of the assigned treatment actually received by the individual. The observed compliance $D_i$ therefore lies in the interval $[0, 1]$.

We will work within the potential outcomes framework [Rubin, 1974] and use a principal stratification approach [Frangakis and Rubin, 2002a] to account for the fact that the observed compliance behavior is observed post-randomization. For individual $i$, we will let $Y_i(0)$ and $Y_i(1)$ represent what the outcome for that individual would be under assignment to $Z_i = 0$ and $Z_i = 1$, respectively. We will let $D_i(0)$ and $D_i(1)$ represent what the compliance behavior for that individual would be under assignment to $Z_i = 0$ and $Z_i = 1$, respectively. We will make the same assumptions about these counterfactual values as in Ma et al. [2011]: SUTVA, consistency, exchangeability,

and the exclusion restriction (as these assumptions were detailed in Chapters 1 and 2 we do not repeat their definitions here).

The causal estimand of interest is the principal causal effect $PCE$ surface, defined as the intent-to-treat effect within strata defined by the potential compliance behaviors $(D(0), D(1))$

$$PCE(d_0, d_1) = E[Y(1)|(D(0), D(1)) = (d_0, d_1)] - E[Y(0)|(D(0), D(1)) = (d_0, d_1)]. \tag{3.1}$$

### *3.3 Identifability Concerns in Partial Compliance Models*

The target of inference in Ma et al. [2011], Jin and Rubin [2008], and Bartolucci and Grilli [2011] is the full $PCE$ surface. However, an investigation of the identifiability of the full $PCE$ surface is missing from these previous works. Concerns about identifiability in partial compliance models are based on the fact that the pairs $(D(0), D(1))$ are never jointly observed. Thus, we must assume a model for the joint distribution of $(D(0), D(1))$, which will depend on an unidentified correlation parameter. Furthermore, as $D(0)$ and $D(1)$ are both continuous, identifying the $PCE$ surface is equivalent to identifying principal strata effects for an infinite number of strata. Given that principal stratification-based approaches often have identifiability issues even with a small number of strata, it is reasonable to ask whether identification of causal effects with an infinite number of strata is possible, even if parameteric assumptions about the $PCE$ surface are used to make the problem of estimation and inference more tractable. In Efron and Feldman [1991], model identifiability follows from the assumption of a deterministic relationship between the potential compliance behaviors. In Jin and Rubin [2008], the authors do not address model identifiability for their Bayesian approach. It is sometimes argued that identifiability is not a concern for Bayesian inference, since inference based on the posterior distribution can proceed regardless of whether the model is identifiable or not. However, as noted in Richardson et al. [2011], unidentified models can lead to posterior distributions that are highly sensitive to the choice of prior. Bartolucci and Grilli [2011] fails to address the basic identifiability of their proposed model. In Ma et al. [2011], the authors acknowledge concerns about the identifiability of the individual parameters in their model for the $PCE$, but conclude that inference for the overall $PCE$ is still possible.

This conclusion is based on low values for mean squared error obtained in the simulation scenarios they considered. However, reliance on simulation results such as these can be misleading. For instance, we can consider the example of having independent, identically distributed observations $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$, and using the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate the variance parameter $\sigma^2$. The MSE in this case is given by $(\mu - \sigma^2)^2 + \frac{\sigma^2}{n}$. In particular cases where $\mu \approx \sigma^2$, the resulting MSE can be quite low. However, this hardly implies that the sample mean is an appropriate estimator of a variance parameter. This example demonstrates the inherent danger in using simulation results to make general confirmative conclusions about statistical properties of an estimator. On the other hand, simulations can be useful for producing counter examples or demonstrating where methods fail.

In the following section, we will consider two approaches to likelihood-based inference for the $PCE$. The first approach is the one used in Ma et al. [2011], which is a two-stage approach based on direct maximization of the likelihood for the regression parameters of the $PCE$ surface after first estimating the parameters for the marginal distributions of the potential compliance behaviors. The second approach, more similar to Bartolucci and Grilli [2011], performs full maximum likelihood estimation using a variant of the Expectation-Maximization (EM) algorithm. In the next section, we describe these models in greater detail. We then demonstrate through simulations that neither the two-stage nor the full maximum likelihood approaches are appropriate for inference about the $PCE$ under two of the most common outcome models, the linear regression model and the logistic regression model.

## *3.4 Likelihood-Based Approaches*

### *3.4.1 Complete-data Likelihood*

We now describe how a likelihood-based approach to modeling partial compliance in RCTs can be carried out. We first consider the complete-data likelihood contribution from a single individual, i.e. the likelihood term for individual $i$ if we were able to observe both $D_i(0)$ and $D_i(1)$. This likelihood will depend on the distribution of the potential outcomes conditional on the potential compliance behaviors as well as the joint distribution of the potential compliance behaviors. We first consider the distribution of the potential outcomes conditional on the potential compliance

behaviors. We assume that the distribution of the potential outcomes comes from an exponential family, so that the density functions $f_0$ and $f_1$ of $Y(0)$ and $Y(1)$, respectively, have the form

$$f_z(y \mid d_0, d_1) = h(y, d_0, d_1)\exp[y\nu_z - A(\nu_z)]$$

for $z = 0, 1$. Using the generalized linear model framework developed by Nelder and Wedderburn [1972], we assume that $\nu_z$ has the form of a linear predictor

$$\nu_z = \boldsymbol{d}^T\boldsymbol{\beta}_z$$

for some $p \times 1$ vector of regression covariates $\boldsymbol{d}$ and a $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}_z = (\beta_{0z}, \beta_{1z}, \ldots, \beta_{pz})^T$. For concreteness and simplicity in notation, we will focus on the case of a regression model with main effects terms only, so that

$$\begin{aligned}\boldsymbol{d}^T &= (1, d_0, d_1)^T \\ &\text{and} \\ \boldsymbol{\beta}_z &= (\beta_{0z}, \beta_{1z}, \beta_{2z})^T.\end{aligned}$$

We note that in the context of partial compliance models, the exclusion restriction assumption implies that $\beta_{00} = \beta_{01}$. The causal estimand $PCE(D(0), D(1))$ can then be expressed as

$$\begin{aligned}E[Y(1) \mid D(0), D(1)] - E[Y(0) \mid D(0), D(1)] &= g^{-1}(\nu_1) - g^{-1}(\nu_0) \\ &= g^{-1}(\beta_{01} + \beta_{11}D(0) + \beta_{21}D(1)) \\ &\quad - g^{-1}(\beta_{00} + \beta_{10}D(0) + \beta_{20}D(1))\end{aligned}$$

for some link function $g$.

As in Ma et al. [2011] and Bartolucci and Grilli [2011], the joint distribution of the potential compliance behaviors $(D(0), D(1))$ will be modeled through a copula. We will model the marginal distributions of $D(0)$ and $D(1)$ using Beta distributions,

$$\begin{aligned}D(0) &\sim \text{Beta}(\alpha_{01}, \alpha_{02}) \\ D(1) &\sim \text{Beta}(\alpha_{11}, \alpha_{12}).\end{aligned} \tag{3.2}$$

Given these marginal distributions $F_{D(0)}(d_0)$ and $F_{D(1)}(d_1)$, we will model the dependence between $D(0)$ and $D(1)$ using a Gaussian copula

$$\begin{aligned} F(d_0, d_1) &= C(\ F_{D(0)}(d_0), F_{D(1)}(d_1)\ ) \\ &\equiv \Phi_2(\ \Phi^{-1}(F_{D(0)}(d_0)),\ \Phi^{-1}(F_{D(1)}(d_1))\ ) \end{aligned} \tag{3.3}$$

where $\Phi_2$ is the distribution function for a bivariate normal random vector with mean $\mathbf{0}$ and correlation matrix $\boldsymbol{R}$ with correlation parameter $\rho$, and $\Phi$ is the distribution function for a standard normal random variable. It can be shown [Fan and Patton, 2014] that the probability density function of $(D(0), D(1))$ is given by

$$p(d_0, d_1) = c_\rho(d_0, d_1) f_{D(0)}(d_0;\ \boldsymbol{\alpha_0}) f_{D(1)}(d_1;\ \boldsymbol{\alpha_1}) \tag{3.4}$$

where $c_\rho$ is the copula density function for the Gaussian copula and $\boldsymbol{\alpha_z} = (\alpha_{z0},\ \alpha_{z1})$ is the vector of parameters that determines the marginal distribution of $D(z)$. As $D(0)$ and $D(1)$ are never jointly observed, the correlation parameter $\rho$ is unidentified.

Letting $\theta = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}, \rho)$ represent the full vector of model parameters, the complete-data likelihood contribution from a single individual in treatment arm $Z_i = z$ is given by

$$L(\theta | Y_i(z), D_i(0), D_i(1)) = f_z(Y_i(z) |\ D_i(0), D_i(1)) c_\rho(D_i(0), D_i(1)) f_{D(0)}(D_i(0)) f_{D(1)}(D_i(1)) \tag{3.5}$$

The observed-data likelihood $L^*$ is then given by integrating the complete-data likelihood with respect to the missing compliance behavior. For either treatment arm, the missing compliance behavior is given by $D_i(1-z)$, so the observed likelihood can be written as:

$$L^*(\theta | Y_i(z), D_i(z)) = \int_0^1 f(Y_i(z) |\ D_i(z), d_{1-z}) p(D_i(z), d_{1-z})\ \mathrm{d}d_{1-z}, \tag{3.6}$$

where it is understood that $p(D_i(z), d_{1-z}) = f(D_i(z), d_{1-z})$ if $z = 0$ and $p(D_i(z), d_{1-z}) = f(d_{1-z}, D_i(z))$ if $z = 1$.

The two-stage approach to estimation and inference as described in Ma et al. [2011] is carried out as follows. First, maximum-likelihood estimates $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\alpha}}_1$ of

the marginal compliance parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are found based on the observed compliance behaviors in the control and treatment arms, respectively. These values can then be substituted into (3.6), and the integral in the observed likelihood can then be estimated by means of a quadrature method, yielding an approximate observed likelihood which is maximized with respect to $\boldsymbol{\beta}$ directly to give an estimate $\hat{\boldsymbol{\beta}}$.

Alternatively, we can work with the complete-data likelihood and use the EM algorithm [Dempster et al., 1977] to obtain parameter estimates. We now detail how the Expectation step (E-step) and Maximization step (M-step) can be carried out for this model.

### *3.4.2 E-Step*

The complete-data log-likelihood is given by

$$\begin{aligned} l(\theta|Y(z), D(z), D(1-z)) &= \log\, f(Y(z)|D(z), D(1-z)) + \log\, f_{D(z),D(1-z)}(D(z), D(1-z)) \\ &= \log\, f(Y(z)|D(z), D(1-z)) + \log\, c_\rho(D(z), D(1-z)) \\ &\quad + \log\, f_{D(z)}(D(z)) + \log\, f_{D(1-z)}(D(1-z)) \end{aligned} \tag{3.7}$$

The Expectation step (E-step) of the EM algorithm consists of finding the expected value of the complete-data log-likelihood, conditional on the observed data and the current estimate $\theta^t$ of the parameter $\theta$. The observed quantities in the log-likelihood are $Y(z)$ and $D(z)$, so the contribution of an individual assigned to treatment arm $Z = z$ to the E-step, $Q_i(\theta|\theta^t)$ is given by

$$\begin{aligned} Q_i(\theta|\theta^t) &= E[l(\theta)|Y_i(z) = y, D_i(z) = d, \theta^t] \\ &= \int_0^1 l(\theta) f(d_{1-z}|y, d, \theta^t)\, \mathrm{d}d_{1-z} \end{aligned} \tag{3.8}$$

Because the regression parameters $(\boldsymbol{\beta_0},\ \boldsymbol{\beta_1})$ and the marginal compliance parameters $(\boldsymbol{\alpha_0},\ \boldsymbol{\alpha_1})$ appear in additively separable terms of the log-likelihood, the linearity of expectation allows us to express $Q_i$ as a sum of two terms, one depending only on the

regression parameters and the other depending only on the compliance parameters

$$\begin{aligned} Q_i(\theta|\theta^t) &= \int_0^1 l(\theta) f(d_{1-z}|y,d,\theta^t)\ \mathrm{d}d_{1-z} \\ &= \int_0^1 \log\ f(y|d,d_{1-z};\boldsymbol{\beta_z})\ f(d_{1-z}|y,d,\theta^t)\ \mathrm{d}d_{1-z} + \int_0^1 \log\ c_\rho(d,d_{1-z}) f(d_{1-z}|y,d,\theta^t)\ \mathrm{d}d_{1-z} \\ &\quad + \int_0^1 \log\ f(d;\boldsymbol{\alpha_z}) f(d_{1-z}|y,d,\theta^t)\ \mathrm{d}d_{1-z} + \int_0^1 \log\ f(d_{1-z};\boldsymbol{\alpha_{1-z}}) f(d_{1-z}|y,d,\theta^t)\ \mathrm{d}d_{1-z} \\ &\equiv Q_i(\boldsymbol{\beta_z}|\theta^t) + Q_i(\boldsymbol{\alpha}|\theta^t) \end{aligned} \tag{3.9}$$

We focus first on evaluating $Q_i(\boldsymbol{\beta_z}|\theta^t)$. This integral generally does not have a closed form solution. Following the approach of Ibrahim and Weisberg [1992] we will approximate it using a quadrature method. The integral can be written

$$\int_0^1 \log\ f(y|d,d_{1-z};\boldsymbol{\beta_z}) f(d_{1-z}|y,d,\theta^t) \approx \sum_{j=1}^{J} a_j \log\ f(y|d,\eta_j) f(\eta_j|y,d,\theta^t) \tag{3.10}$$

for weights $a_j$ and zeros $\eta_j$ depending on the number of nodes $j = 1, 2, \ldots, J$ chosen for the quadrature. The density function of $D(1-z)$ conditional on $Y(z) = z, D(z) = z, \theta^t$ can be written as

$$f(d_{1-z}|y,d,\theta^t) = \frac{f(y,d,d_{1-z})}{f(y,d)} \tag{3.11}$$

Plugging 3.11 into 3.10, we can approximate $Q_i(\boldsymbol{\beta_z}|\theta^t)$ as

$$\begin{aligned} \int_0^1 \log\ f(y|d,d_{1-z};\boldsymbol{\beta_z}) f(d_{1-z}|y,d\theta^t) &\approx \sum_{j=1}^{J} a_j \log\ f(y|d,\eta_j) f(\eta_j|y,d,\theta^t) \\ &= \sum_{j=1}^{J} w_j f(y|d,\eta_j) \end{aligned} \tag{3.12}$$

where the weights $w_j$ are given by

$$w_j = \frac{f(y|d,\eta_j;\theta^t) f(d,\eta j)}{f(y,d;\theta^t)} \tag{3.13}$$

Evaluating $Q_i(\boldsymbol{\alpha}|\theta^t)$ can be done in a similar manner. We then have that

$$\begin{aligned} Q_i(\boldsymbol{\alpha}|\theta^t) &= \int_0^1 \left(\log\, c_\rho(d, d_{1-z}) + \log\, f(d|\boldsymbol{\alpha_z}) + \log\, f(d_{1-z}|\boldsymbol{\alpha_{1-z}})\right) f(d_{1-z}|y, d, \theta^t)\, \mathrm{d}d_{1-z} \\ &\approx \sum_{j=1}^{J} w_j \left(\log\, c_\rho(d, \eta_j) + \log\, f(d|\boldsymbol{\alpha_z}) + \log\, f(\eta_j|\boldsymbol{\alpha_{1-z}})\right) \end{aligned}$$

where the weights $w_j$ are given by 3.13. Given a sample of $n$ independent observations, where individuals are randomized to either receive the treatment $Z = 1$ or the control $Z = 0$, we can write the E-step for the entire sample as

$$\begin{aligned} Q(\theta|\theta^t) &= \sum_{i:Z_i=0} Q_i(\theta|\theta^t) + \sum_{i:Z_i=1} Q_i(\theta|\theta^t) \\ &= \sum_{i:Z_i=0} \left(Q_i(\boldsymbol{\beta_0}|\theta^t) + Q_i(\boldsymbol{\alpha}|\theta^t)\right) + \sum_{i:Z_i=1} \left(Q_i(\boldsymbol{\beta_1}|\theta^t) + Q_i(\boldsymbol{\alpha}|\theta^t)\right) \end{aligned}$$

### *3.4.3 M-step*

We now consider how the Maximization step (M-step) of the EM algorithm can be performed to obtain an updated estimate of the parameter $\theta$. Because the regression parameters $\boldsymbol{\beta}$ and the compliance parameters $\boldsymbol{\alpha}$ appear in separate terms of $Q(\theta|\theta^t)$, we can consider separately how to update them. It can be shown [Ibrahim and Weisberg, 1992] that the updated regression parameters $\boldsymbol{\beta_0}^{t+1}$ and $\boldsymbol{\beta_1}^{t+1}$ can be found by fitting a weighted logistic regression to an augmented version of the observed data. The augmented dataset consists of $J \times n$ observations, where for the $i^{th}$ individual we construct $j = 1, 2, \ldots, J$ observations with $Y_{ij} = Y_i, Z_{ij} = Z_i, D_{ij}^{observed} = D_i,$ and $D_{ij}^{missing} = \eta_j$. The weight assigned to each observation $Y_{ij}$ is given by (3.13).

Maximizing the expected log-likelihood related to the $\boldsymbol{\alpha}$ parameters is complicated by the presence of the copula density $c_\rho$. If the potential compliance behaviors were assumed to be independent, then the updated $\boldsymbol{\alpha}$ parameters could be found as the MLEs from two independent beta distributions. The copula density term in the expected log-likelihood, however, depends on the $\boldsymbol{\alpha}$ parameters in a complicated way due to the fact that the distributions functions $F_{D(0)}$ and $F_{D(1)}$ appear in the copula

density, as shown in 3.14.

$$c_\rho(d_0, d_1) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\mathbf{u}(\mathbf{R}^{-1} - \mathbf{I})\mathbf{u}^T\right\} \tag{3.14}$$

where

$$\mathbf{u} = (\Phi^{-1}(F_{D(0)}(d_0; \boldsymbol{\alpha_0})), \Phi^{-1}(F_{D(1)}(d_1; \boldsymbol{\alpha_1})))$$

The updated potential compliance parameter estimates $\boldsymbol{\alpha}^{t+1}$ can be found by maximizing $Q(\boldsymbol{\alpha}|\theta^t)$ through a numerical optimization procedure.

## *3.5 Simulations Demonstrating Inconsistency of Maximum likelihood-based Partial Compliance Methods*

Having discussed how likelihood-based inference can be performed for partial compliance models through either the two-stage model proposed by Ma et al. [2011] or the weighted EM algorithm, we now present simulations demonstrating consistency issues for estimators based on either of these approaches.

First, we consider the case of a normally distributed outcome $Y_i$. The potential outcomes $Y_i(z), z = 0, 1$ conditional on $D_i(0) = d_0$ and $D_i(1) = d_0$ were simulated from the following distribution:

$$Y_i(z) \sim N(\beta_0 + \beta_{1z} D_i(0) + \beta_{2z} D_i(1), \sigma^2),$$

with $\boldsymbol{\beta}_0 = (\beta_0, \beta_{10}, \beta_{20}, \beta_{11}, \beta_{21}) = (1.3, 2.1, 1.4, -.5, 2.5)$ and $\sigma^2 = 1$. The correlation between potential outcomes $Y_i(0)$ and $Y_i(1)$ is set to 0.3. The potential versions of treatment were simulated from a Gaussian copula model with marginal $Beta(3, 2)$ distributions and correlation parameter $\rho = 0.4$.

Using these settings, we generated datasets at sample sizes of $n = 1000, 2000, \ldots, 100000$. At each sample size, we estimated the regression parameters $\boldsymbol{\beta}$ using both the two-stage maximum likelihood approach and the weighted EM algorithm approach. In fitting both methods to the data, we used the true value for $\rho$ as our working correlation value and used a Gaussian copula for modeling the joint distribution of the potential versions of treatment. In addition, we fit a regression model matching the true underlying regression model. Hence, there was no model misspecification. In Figure 3.1, we show traceplots of the regression parameter estimates from the two-

stage maximum-likelihood and the weighted EM algorithm methods. It is clear from this figure that the two-stage maximum-likelihood method is producing inconsistent estimates for the individual regression parameters, even under the best-case scenario of no misspecification of both the model for the potential versions of treatment and the model for the potential outcomes. In Ma et al. [2011], the authors note that their simulations also indicate an inability to estimate the individual regression parameters; however, they argue that accurate estimation of the overall $PCE$ surface is still possible. Figure 3.2 shows a $3x3$ array of trace plots for estimates of selected points on the $PCE$ surface. From this figure, we can see that the two-stage maximum likelihood procedure produces inconsistent estimation of the $PCE$ surface in addition to inconsistent estimation of the individual regression parameters. On the other hand, the weighted EM algorithm appears to produce consistent estimates of the regression parameters and the points on the $PCE$ surface. As we noted before, however, it is dangerous to make affirmative conclusions regarding consistency of estimators based solely on results from specific simulations. Furthermore, in the next section we present simulations under misspecification of the correlation parameter $\rho$ demonstrating that estimation of the $PCE$ surface based on the weighted EM algorithm is likely to lead to misleading results in practice, even if the weighted EM algorithm produces consistent estimates for a linear model under the best-case scenario of no model misspecification.

We now present simulation results for the case of a binary outcome $Y_i$. The model for the potential outcome $Y_i(z)$ conditional on $D_i(0) = d_0$ and $D_i(1) = d_1$ is given by the following logistic regression model:

$$\log\frac{(P(Y_i(z)=1)}{1-P(Y_i(z)=1)} = \beta_0 + \beta_{1z}D_i(0) + \beta_{2z}D_i(1),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_{10}, \beta_{20}, \beta_{11}, \beta_{21}) = (1.3, 2.1, 1.4, -.5, 2.5)$. As in the previous simulations, the distribution of the potential versions of treatment was based on a Gaussian copula model with correlation parameter $\rho = 0.4$ and marginal distributions given by a $Beta(3, 2)$ distribution. We generated datasets at sample sizes of $n = 1000, 2000, 3000, \ldots, 100000$. For each dataset, we used the two-stage maximum likelihood and weighted EM algorithm methods to produce estimates of the regression parameters. Figure 3.3 shows trace plots for the estimates of the individual

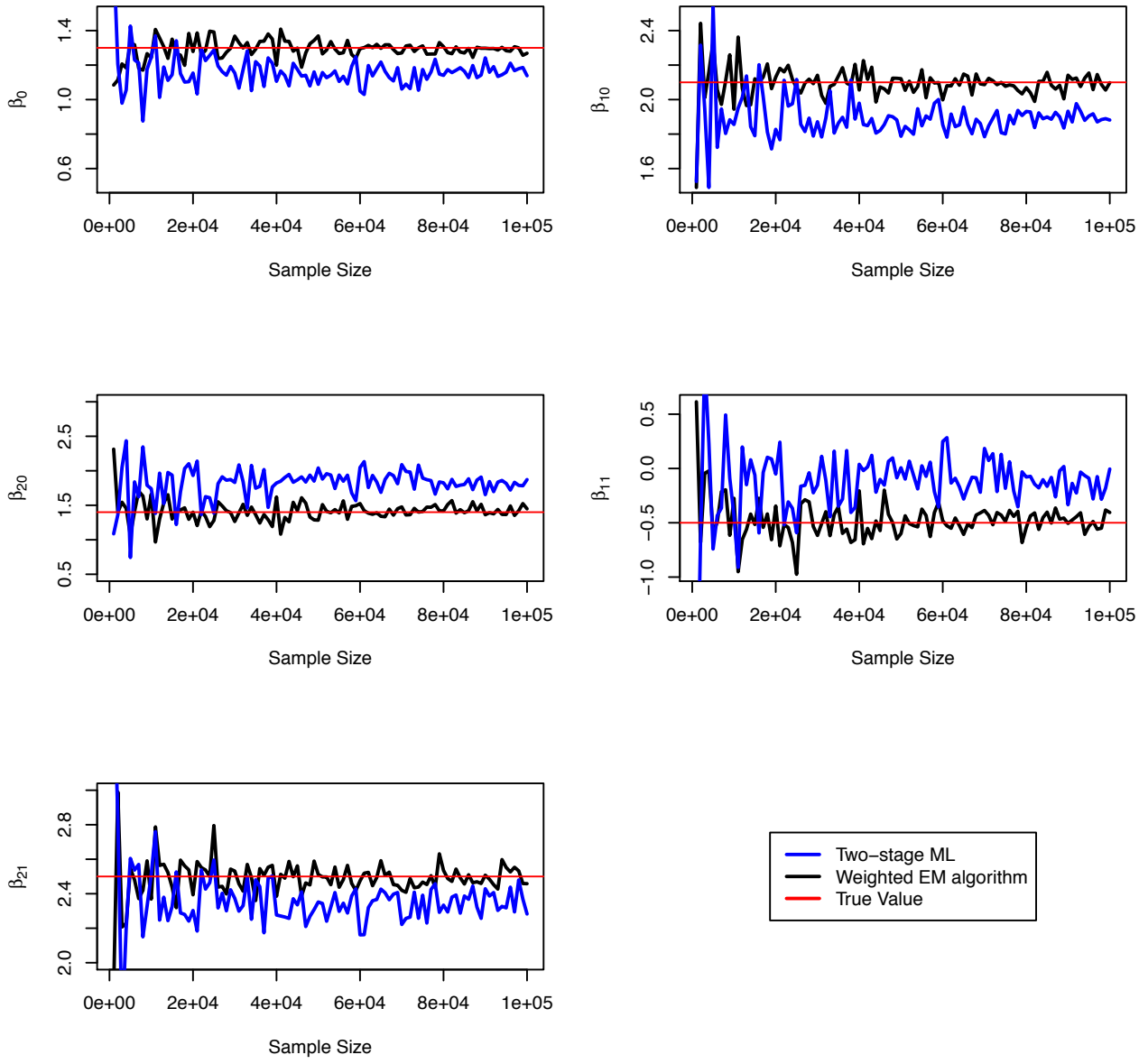



Figure 3.1: Trace plots showing linear regression parameter estimates from the two-stage maximum-likelihood approach and the weighted EM algorithm approach as a function of increasing sample size. In each panel, the blue line shows the two-stage estimates, the black line shows the weighted EM estimates, and the red line shows the true underlying value for the regression parameter.

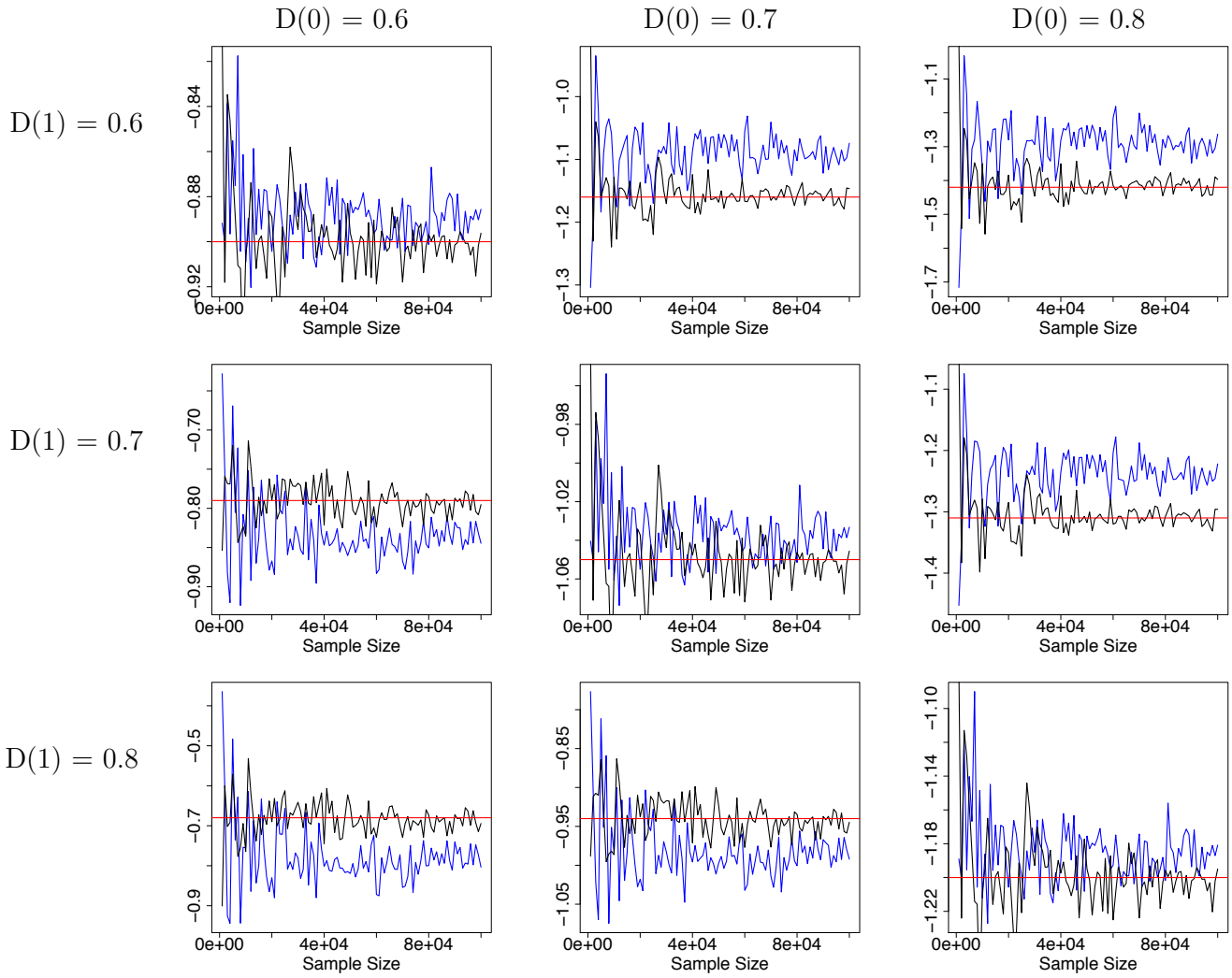



Figure 3.2: Array of trace plots showing estimates of selected points on the $PCE$ surface from a linear outcome model as a function of increasing sample size. Each cell in the $3x3$ array shows estimates for $PCE(D(0), D(1))$, e.g. the upper left-hand cell shows estimates for $PCE(0.6, 0.6)$. In each cell, the blue line shows estimates from the two-stage maximum-likelihood approach, the black line shows estimates from the weighted EM algorithm, and the red line shows the true value for $PCE(D(0), D(1))$. In each cell, the x-axis ranges from $n = 1000$ to $n = 100000$.

regression parameters as a function of increasing sample size, while Figure 3.4 shows similar trace plots for estimates of selected points on the $PCE$ surface.

From these figures, we can see that the consistency issues present in the case of a linear model only become worse as we move to a model with a non-linear link function. The weighted EM algorithm is now clearly producing inconsistent estimates of both the individual regression parameters and certain points on the $PCE$ surface. The estimates of the regression parameters from the two-stage maximum likelihood approach, on the other hand, still exhibit such a degree of variability even at sample sizes as large as $n = 100000$ that it is not clear what, if any, value they are converging to. Inspection of the traceplots for the $PCE$ estimates shows that the two-stage maximum likelihood approach produces either inconsistent or highly variable estimates of points on the $PCE$ surface generated by this logistic regression model.

## *3.6 Simulations Demonstrating Bias under Misspecification of Correlation between Potential Compliance Behaviors*

In the previous section, we conducted simulations assuming that the true value of $\rho$, the correlation between $D(0)$ and $D(1)$, was known. These simulations demonstrated that the two-stage maximum likelihood procedure is inappropriate for both a linear or logistic model for the potential outcomes, even with no model misspecification. On the other hand, the weighted EM algorithm, while clearly inconsistent even with no model misspecification under a logistic model, appears to yield consistent estimators of the regression parameters for a linear outcome model when there is no model misspecification. However, we now demonstrate that the regression parameter estimates (and by extension the estimates of the $PCE$ surface) from the weighted EM algorithm exhibit non-neglible bias when the correlation parameter $\rho$ is misspecified.

To demonstrate the bias introduced by misspecification of the correlation parameter $\rho$, we performed simulations using the same data-generating mechanism for a linear outcome model described in the previous section. Under this setup, we generated 1000 datasets with a sample size of $n = 1000$ However, we now allowed for misspecification of the correlation parameter $\rho$ when fitting the weighted EM algorithm model to the data. We used assumed values of $\rho$ ranging from 0.1 to 0.7 in increments of 0.1.

Figure 3.5 shows the average values for the regression parameters across the

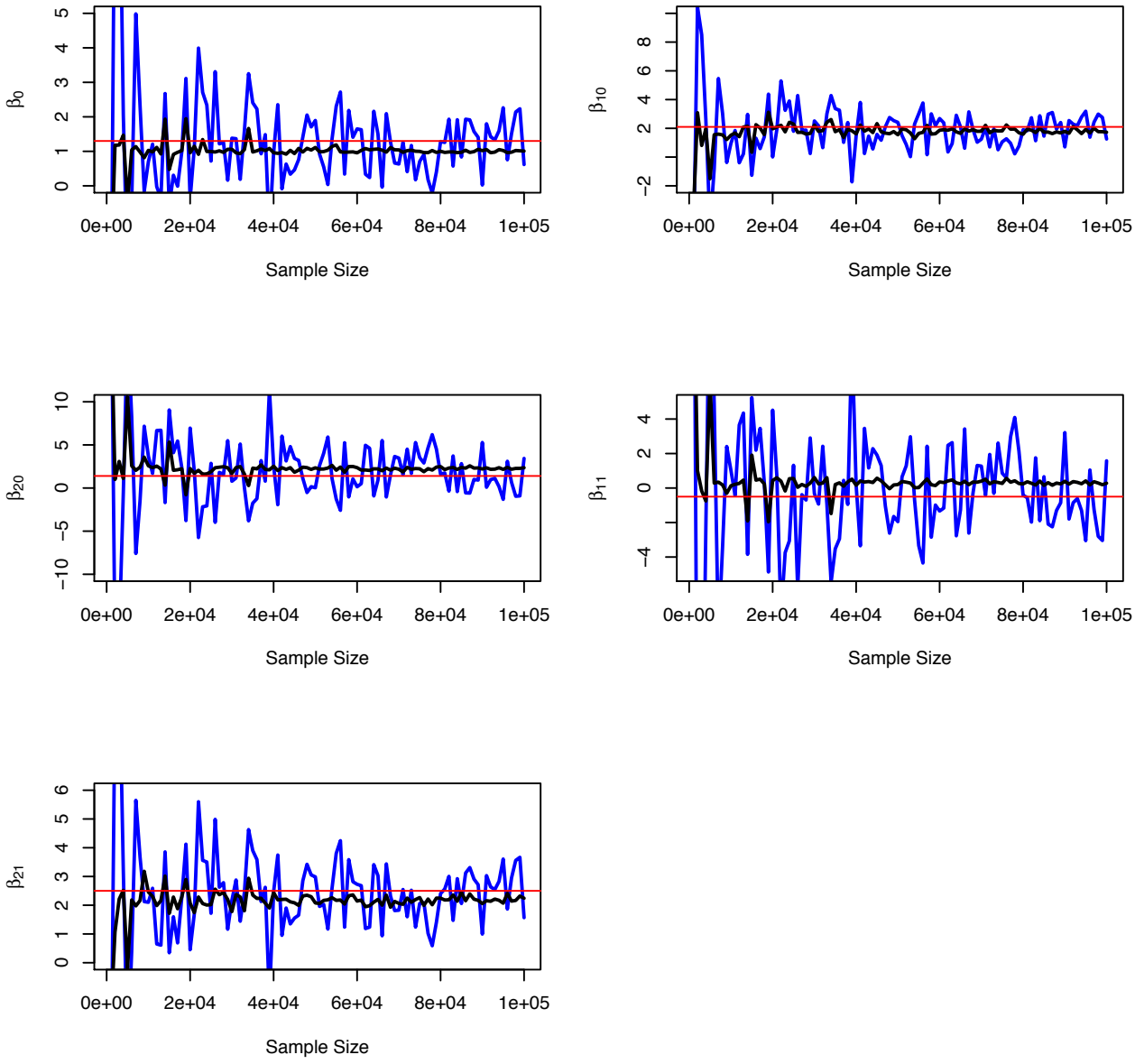



Figure 3.3: Trace plots showing logistic regression parameter estimates from the two-stage maximum-likelihood approach and the weighted EM algorithm approach as a function of increasing sample size. In each panel, the blue line shows the two-stage estimates, the black line shows the weighted EM estimates, and the red line shows the true underlying value for the regression parameter.

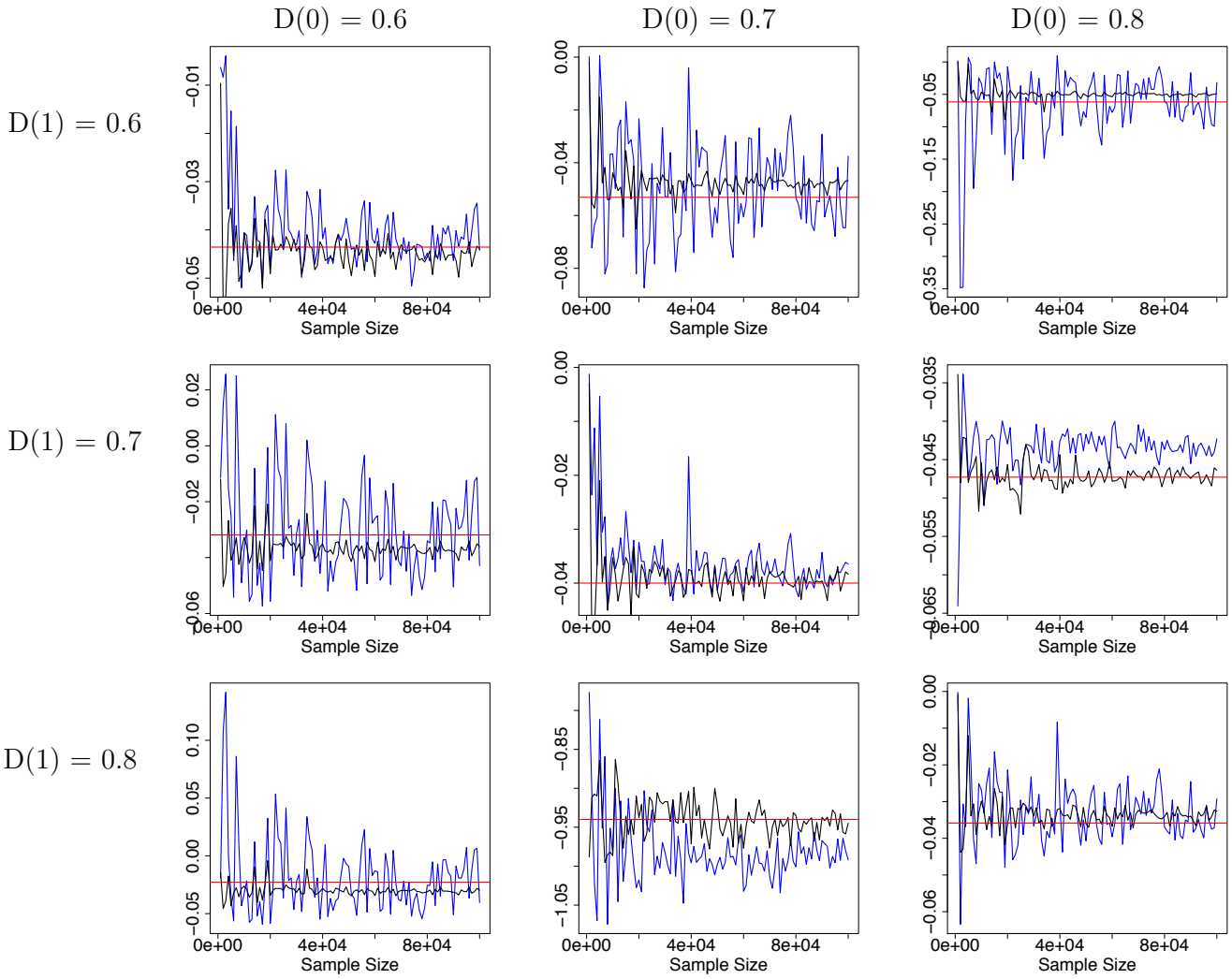



Figure 3.4: Array of trace plots showing estimates of selected points on the $PCE$ surface from a logistic outcome model as a function of increasing sample size. Each cell in the $3x3$ array shows estimates for $PCE(D(0), D(1))$, e.g. the upper left-hand cell shows estimates for $PCE(0.6, 0.6)$. In each cell, the blue line shows estimates from the two-stage maximum-likelihood approach, the black line shows estimates from the weighted EM algorithm, and the red line shows the true value for $PCE(D(0), D(1))$. In each cell, the x-axis ranges from $n = 1000$ to $n = 100000$.

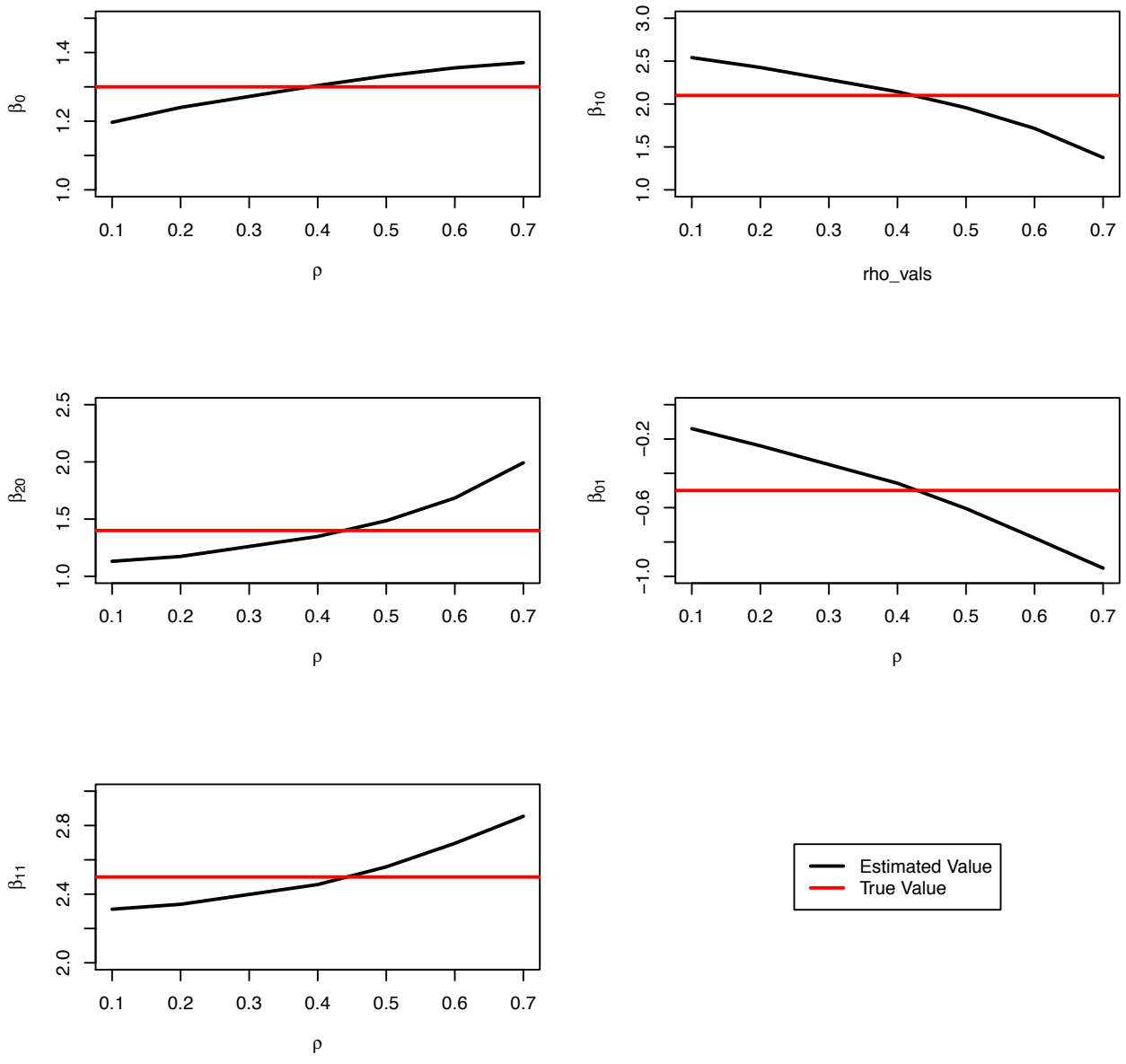



Figure 3.5: Mean values for the regression parameter estimates from the weighted EM algorithm as a function of the assumed value for $\rho$ used in fitting the model. In each panel, the red line shows the true regression parameter value used in generating the data.

1000 simulations plotted against the assumed value for $\rho$ that was used in fitting the model. When $\rho = 0.4$ (i.e. the model is correctly specified), the regression parameter estimates are on average fairly close to the true values. However, when the value of the correlation parameter is misspecified, there is clear bias introduced in the estimation of each of the regression parameters. As the assumed value of $\rho$ moves further from the true value in either direction, the magnitude of the bias becomes greater. In practice, of course, we have no way of learning about the true value of $\rho$ from the data, and it is extremely unlikely that we would actually specify the true value for this correlation parameter. Thus, even if we are willing to take the simulations from the previous section as definitive and conclude that the weighted EM algorithm gives consistent estimates of the regression parameters from a linear model under no model misspecification, we are still left with the fact that in practice we will have biased and inconsistent estimates due to the near certainty of misspecification of $\rho$. While the degree of bias shown in these simulations is not particularly severe, we speculate that this is likely due to the choice of model parameters for these specific simulations. For other choices of the underlying model parameters, the degree of bias may be quite severe.

## *3.7 Discussion*

In this chapter, we have demonstrated that the approach to causal inference in RCTs with partial compliance proposed by Ma et al. [2011] produces clearly inconsistent estimates of the $PCE$ surface under a linear outcome model. Furthermore, under a logistic regression model for a binary outcome, their method produces estimates whose consistency or inconsistency is not clear due to continued variability even at sample sizes as large as $n = 100000$. Since two-stage maximum likelihood approaches are known to have other issues [Murphy and Topel, 2002], one might ask whether these consistency problems can be solved by using a full maximum likelihood approach. However, we have additionally demonstrated that a full maximum likelihood-based approach using a weighted EM algorithm leads to clearly inconsistent estimation of the $PCE$ surface when the outcome variable is binary and a logistic regression model is assumed. All of these results demonstrating the deficiency of these models were produced in the best-case scenario assuming no misspecification of either the model for the potential outcomes or the model for the potential compliance behaviors. The fact

that these methods fail for common outcome models even in these best-case scenarios does not bode well for their utility when applied to real-world data sets where some degree of model misspecification is nearly guaranteed, particularly misspecification of the unidentified correlation between the potential compliance behaviors. Furthermore, we have demonstrated that misspecification of this correlation parameter leads directly to biased estimates of the regression parameters and by extension the $PCE$ surface.

For many interventions, compliance behavior is more naturally described in terms of the proportion of the assigned treatment that was received. In pharmaceutical trials, the assigned treatment may in fact be a treatment regimen spanning several months. For trials of behavioral interventions, the intensity of the treatment as measured by the number of hours spent in therapy may be of interest. Applying the standard all-or-none non-compliance methods to these RCTs requires deciding how to dichotomize these continuous measures of compliance behavior. Hence, the desire for methods that treat compliance behavior in RCTs as a continuous rather than binary variable is understandable. However, just because something is desired does not mean it can be obtained. Given the deficiencies of these maximum-likelihood based methods as well as their total reliance on the unidentified joint distribution of the potential versions of treatment, we must conclude that the use of these models is more likely than not to lead to unjustifiable and potentially misleading conclusions.

# Chapter 4

# ANALYZING RANDOMIZED TRIALS WITH NON-COMPLIANCE USING THE `NONCOMPLYR` R PACKAGE

## *4.1 Preface*

In this chapter, we present a tutorial on Bayesian methods for performing causal inference with data from RCTs in the presence of binary non-compliance. The tutorial can be seen as a companion piece to `noncomplyR`, a software package for the R statistical programming language that I developed as a way for researchers to easily implement the methods for causal inference in RCTs with noncompliance described in Imbens and Rubin [1997]. The `noncomplyR` package is publicly available from the Comprehensive R Archive Network (https://cran.r-project.org). The tutorial gives a conceptual overview of causal inference using potential outcomes in the presence of non-compliance to treatment assignment, and then gives detailed explanations of how compliance-based methods can be implemented with `noncomplyR`.

Unlike the previous chapters, which were written with a statistically and mathematically sophisticated audience in mind, this chapter has been written with the intended audience of the scientist with a basic statistical education. Thus, this target audience should be kept in mind when reading this tutorial. Researchers such as these comprise a significant portion of the people performing statistical analysis in the real world. For these researchers, there are two main barriers to implementing more advanced methods such as the ones presented in this tutorial. First, these methods are often not a part of their statistical training, and they may lack the mathematical background necessary to learn about these methods on their own from statistical research papers. Second, ready-to-use statistical software for these more advanced methods may not be available. This tutorial addresses both of these issues by explaining the concepts in an approachable way and by providing the necessary software tools to immediately implement the methods with real data. Tutorials such as this one, tai-

lored to a specific audience and published in non-statistical subject matter journals that scientists and researchers read, can do a great service in improving the statistical sophistication of the scientific world.

### *4.2 Introduction*

Randomized clinical trials (RCTs) are considered the gold standard for assessing the efficacy of behavioral interventions. Randomization protects against confounding, and as a result RCTs allow researchers to determine causal effects of interventions. Non-adherence to the intervention protocol, a frequently encountered issue in RCTs, complicates the analysis of data from RCTs. In this tutorial, we show how information about individual-level adherence to treatment assignment can be incorporated into the analysis of data from an RCT in a statistically rigorous way. We begin by discussing limitations to three common approaches to analyzing data from RCTs in the presence of non-adherence to the intervention protocol. We argue that, when non-adherence is present, these common analysis approaches may inadequately estimate the causal effect of actually receiving the intervention. We then give a conceptual overview of an approach to estimating the causal effect of receiving the intervention that can overcome these limitations in the presence of non-adherence. We then give a detailed explanation of how this approach can be applied to real data using the R package noncomplyR. We note that the reasons for non-adherence are multiple, encompassing factors both inside and outside the control of study participants and researchers (e.g., research design specifying hours of treatment delivery vs. child illness or weather events). Hence, the term non-adherence is not meant to imply that the individuals themselves have made a conscious decision to not adhere to their treatment assignment. The most common approach to analyzing data from RCTs is the intention-to-treat (ITT) analysis. In an ITT analysis, comparisons are made between groups defined solely by intervention assignment without taking adherence to intervention assignment into account. An ITT analysis gives a valid estimate of the causal effect of assignment to intervention. This effect is often of interest to policy-makers, since it resembles the expected effect of recommending an intervention for use to the population of interest. In the case of perfect adherence to intervention assignment, an ITT analysis also provides a valid estimate of the effect of receiving the treatment with a causal interpretation. However, when non-adherence to

intervention assignment is present, an ITT analysis no longer estimates the effect of receiving the intervention, since some individuals assigned to intervention may not have actually received the intervention, or may have received a lower intensity or modified form of the intervention that potentially violates the intervention protocol. In addition, it is also possible that individuals assigned to the control condition may have received the active intervention. As a result, comparison groups based on intervention assignment are no longer synonymous with the treatment actually received. To address this shortcoming of the ITT analysis when non-adherence has occurred, various alternative analytic approaches have been used. One such alternative is to perform an As-Treated analysis, which makes comparisons between groups defined by the intervention actually received, rather than the original intervention assignment. This approach is limited by the fact that the amount or intensity of intervention is observed at the end of the trial, after randomization has occurred. If adherence to the intervention protocol is perfect, then the As-Treated analysis is equivalent to the ITT analysis. However, if there is non-adherence present in the data, then the comparison groups in an As-Treated analysis will not be randomized. If adherence behavior is related to factors that influence intervention response and outcomes, then this would introduce confounding factors. There may be imbalances between the comparison groups in key variables related to the things that determine intervention protocol adherence, such as the number of intervention hours received. As a result, this approach discards one of the main strengths of an RCT (protection against confounding) and significantly limits the ability to interpret the estimated intervention effect in a causal manner. Another alternative analysis approach is a Per-Protocol analysis. A Per-Protocol analysis discards data on individuals who did not follow the study protocol. In contrast to the As-Treated Analysis, this approach does not reorganize the comparison groups based on participant behavior post-randomization. However, it does introduce non-random factors into the sample composition that will affect the resulting estimate of the effect of intervention. For example, if individuals that do not follow the intervention protocol are those for whom the protocol is ineffective, the treatment effect will be overestimated. Alternatively, if an intervention is most effective for low-resource families, but those families tend to have lower adherence due to transportation issues, the treatment effect could be under-estimated. Thus, this type of analysis is likely to give biased estimates of the effect of intervention. In

summary, each of the approaches outlined above has limitations for addressing the question of the causal effect of actually receiving an intervention when non-adherence to treatment assignment is present. The ITT analysis ignores non-adherence, and therefore does not estimate the effect of actually receiving the intervention. The As-Treated analysis takes information on non-adherence into account in the analysis, but does so in such a way that a causal interpretation of the results is no longer valid. The Per-Protocol analysis introduces bias by removing a subset of the sample non-randomly. All approaches reduce or eliminate the strength of the RCT design and the ability of investigators to answer questions related to the causal effect of receiving an intervention when non-adherence has occurred. We now outline an approach that overcomes many of these limitations and allows us to make stronger causal inferences in the presence of non-adherence.

### *4.3 Causal Inference, Potential Outcomes and Adherence*

We begin by introducing the basic statistical framework for causal inference that will be used to develop the adherence-based causal analysis approach. This framework is based on the idea of potential outcomes, also known as counterfactuals, and is sometimes referred to as the Rubin causal model [Rubin, 1974]. We suppose that we have a sample of N individuals taking part in a randomized controlled trial. We let Z denote assignment to either the control group (Z = 0) or the active intervention group (Z = 1) and Y denote the outcome of interest. For example, Y may be the score of an assessment test administered at a follow-up visit. In addition to the observed outcome Y, we will assume that each individual has potential outcomes, a pair of values (Y (0), Y (1)) that represents an individuals hypothetical outcomes if the person were assigned to the control group (Y (0)) or active intervention (Y (1)). We express causal effects of treatment in terms of the potential outcomes because (Y (0), Y (1)) represent the responses of an individual under circumstances that are identical except for the intervention assignment. The fundamental problem of causal inference with potential outcomes is that we can only observe at most one of the two potential outcomes for any individual. If an individual is assigned to the active intervention, for example, the observed outcome Y for that individual is equal to that individuals potential outcome under assignment to active intervention, Y(1). The potential outcome under assignment to control, Y(0), will consequently be

missing for that individual. Individual-level causal effects would require knowledge of both potential outcomes, which are not observable except under unusual or special circumstances such as the intervention having the same effect in all individuals. Therefore, individual-level causal effects will typically be unidentified. However, we are often interested in quantities such as averages that describe characteristics of a population, rather than individual-level causal effects. This, combined with the statistical properties that result from randomization of intervention assignment, allows us to estimate average causal effects of treatment by making comparisons between groups of individuals in the study. The potential outcomes framework rests on some assumptions that must hold in order for the potential outcomes, as defined above, to be well defined. The first assumption is known as the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). This assumption states that an individuals potential outcomes depend only on his or her treatment assignment, not on the treatment assignment of any other individuals in the study. This is often called the non-interference assumption. The second assumption, sometimes considered a part of SUTVA, is known as the consistency assumption [VanderWeele, 2009]. This assumption states that if an individual is assigned to treatment group Z = z, then their observed outcome is equal to their potential outcome under that treatment group assignment. This gives us the following relationship between the observed outcome Y and the potential outcomes

$$Y = Y(z)$$

The non-interference and consistency assumptions will be assumed to hold in all of the examples that follow. We will also have reason to consider two other assumptions. Although they are not necessary for the basic potential outcomes framework outlined above, they will play an important role in the estimation of adherence-based causal estimates. These assumptions are known as the exclusion restriction and the strong-access monotonicity assumption. The exclusion restriction assumption states that the treatment assignment has no effect on the outcome except through the treatment actually received. The exclusion restriction assumption is typically hard to justify in observational studies but is often reasonable in RCTs, especially when treatment assignment is blinded. The strong-access monotonicity assumption states that indi-

viduals assigned to the control group have no way of receiving the active intervention. This assumption is most likely to hold in RCTs where the active intervention is under the direct control of the investigators. To see how causal effects defined in terms of potential outcomes can be estimated using the observed data, we can consider the ITT analysis described above in a study where every individual adheres to his or her treatment assignment. Recall that this analysis makes comparisons between those assigned to active intervention and those assigned to control. From randomization of treatment assignment and the consistency assumption, we can say that the average observed outcome within the active intervention group is equal to the average potential outcome under assignment to the active intervention. Similarly, the average observed outcome within the control group is equal to the average potential outcome under assignment to control. Therefore, the comparison in outcomes between the active intervention and control groups gives us an estimate of the causal effect of intervention assignment at the population level. Letting $E$ designate the expectation of a random variable, we have more formally that

$$\begin{aligned} E[Y(1) - Y(0)] &= E[Y(1)] - E[Y(0)] \\ &= E[Y(1) \mid Z = 1] - E[Y(0) \mid Z = 0] \\ &= E[Y \mid Z = 1] - E[Y \mid Z = 0] \end{aligned} \tag{4.1}$$

where the second equality holds through randomization, and the third equality holds from the consistency assumption. This example demonstrates one of the main elements of a causal analysis: expressing the causal effect of interest (in this case $E[Y(1)] - E[Y(0)]$) in terms of quantities found in the observed data (in this case $E[Y \mid Z = 1] - E[Y \mid Z = 0]$). We now consider the scenario where some individuals in the study do not adhere to the intervention to which they are assigned. The treatment assignment, denoted by Z, is assumed to be under the control of the investigators and randomized. However, the intervention actually received by an individual, denoted by D, may not be under the control of the investigators and may in fact differ from the treatment that individual was assigned to receive. This is referred to as non-adherence to treatment assignment. We let D = 0 indicate that the individual received the control or did not receive the active intervention, according to how the control is defined in the context of the study. Similarly, D = 1 indicates that

the individual received the active intervention, according to how receiving the active intervention is defined in the context of the study. If an individual adheres to their treatment assignment, then their value for Z will equal their value for D. However, when an individual does not adhere to the study protocol then D will not equal Z for that individual. When non-adherence is present in the data, the question becomes how to take information about adherence into account while still preserving a causal interpretation of the effect of receiving the intervention. The observed adherence D occurs post-randomization. Consequently, results from an analysis that simply conditions on D cannot be interpreted causally, due to the fact that comparison groups based on D are no longer randomized. A technique in causal inference known as principal stratification offers a general framework for adjusting for post-randomization variables while maintaining a causal interpretation of the results [Frangakis and Rubin, 2002a]. We now outline how the principal stratification framework is applied to the setting of an RCT with non-adherence. We begin by extending the idea of a potential outcome (the outcome that would be observed under assignment to active intervention or control) to the idea of a potential adherence behavior. An individuals potential adherence behavior reflects their adherence to the study protocol that would be observed under assignment to active treatment or that observed under assignment to control. Similar to the potential outcomes defined for the outcome of interest Y, we can think of a pair of values (D(0), D(1)) representing an individuals adherence to treatment assignment had they been assigned to the control group (D(0)) or the active treatment group (D(1)). If the pair of values (D(0), D(1)) were known for each individual, the potential adherence values could be used to classify individuals into subgroups based on adherence behavior, as shown in Table 1. Individuals with (D(0),D(1)) = (1,1), the Always Taker adherence subgroup, are individuals who will always receive the treatment, regardless of which treatment arm they are assigned to. Individuals with (D(0),D(1)) = (0,0), the Never Taker adherence subgroup, are individuals who will never receive the active intervention, regardless of which treatment arm they are assigned to. Individuals with (D(0),D(1)) = (1,0), the Defier adherence subgroup, are individuals who will receive the active intervention when assigned to the control arm but will not receive the active intervention when assigned to the active intervention arm. Finally, individuals with (D(0),D(1)) = (0,1) represent individuals who will not receive the active intervention when assigned to the control arm but

| | $D(1)=0$ | $D(1)=1$ |
|---|---|---|
| $D(0)=0$ | Never Taker | Complier |
| $D(0)=1$ | Defier | Always Taker |

Table 4.1: Possible compliance types in the binary non-compliance framework.

will receive active intervention when assigned to the active intervention arm. This subgroup is known as the Complier subgroup, and will be particularly important in the analysis approaches that follow.

Of the four adherence subgroups, only the Complier and Defier subgroups provide us with information about the effect of actually receiving treatment. This is because only the individuals in the Complier and Defier groups will actually receive or not receive the intervention depending on which intervention arm they are assigned to. The Always Taker and Never Taker groups, on the other hand, do not differ between the control and active intervention arms in terms of what intervention they receive. Individuals in the Always Taker group will always receive the active intervention regardless of which intervention arm they are assigned to, while the Never Takers will never receive the intervention regardless of which intervention arm they are assigned to. For both of these adherence subgroups, any differences between those assigned to the control condition and those assigned to the active intervention would be due to the effect of the intervention assignment itself. Thus, an estimate of the causal effect of actually receiving treatment can be obtained by making comparisons between the average potential outcomes among individuals in either the Complier or Defier adherence subgroups. It is common in adherence-based analyses to make the assumption that no individuals fall into the Defier adherence subgroup. We will make this assumption as well, so that all of the information about the casual effect of actually receiving the treatment is contained within the Complier adherence subgroup. This causal effect is referred to as the Complier Average Causal Effect (CACE)

$$CACE = E[Y(1) \mid \text{ Complier}] - E[Y(0) \mid \text{ Complier }] \tag{4.2}$$

Unlike the causal effect defined in Equation 4.1, the CACE is not as easily calcu-

lated from the observed data due to the fact that the adherence subgroups are not, in general, observed for every individual. We now outline how this difficulty can be overcome using a Bayesian modeling approach. Bayesian Approach to Estimation of the CACE In a Bayesian analysis, a probability model for the data that depends on one or more unknown parameters is first specified. The parameters that determine the form of the probability model (e.g. the mean and variance from a normal distribution, or the probability of a success for a dichotomous outcome) are then given prior distributions. The prior distributions represent the investigators prior knowledge or belief. The choice of priors can range considerably, from very precise (for instance, if results from a large number of previous studies are available) to non-informative (e.g. when investigating a previously unstudied intervention). The key step in a Bayesian analysis is finding the posterior distribution of the model parameters, which can be thought of as the result of updating the prior distribution based on the information contained in the data. Formally, the prior distribution $\pi(\theta)$ and the data model $f(y \mid \theta)$ can be combined by Bayes Theorem to find the posterior distribution

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta) \mathrm{d}\theta}$$

Estimation and inference in a Bayesian analysis are then based on quantities computed from the posterior distribution. A common approach is to generate a sample of parameter values from the posterior distribution. Quantities of interest can then be directly computed based on this sample. We now present a Bayesian framework for adherence-based analyses of data from RCTs. This framework is based on the models described in Imbens and Rubin [1997]. In addition to the observed variables Y, Z, and D, we will let C indicate the adherence subgroup that an individual belongs to, where C = c indicates that an individual belongs to the Complier subgroup, C = n indicates that an individual belongs to the Never Taker subgroup, and C = a indicates that an individual belongs to the Always Taker subgroup. Our approach will be to take advantage of the fact that estimation of the CACE would be straightforward if each

individuals adherence group C were actually observed. In that case, we would have

$$\begin{aligned} CACE &= E[Y(1) \mid C = c] - E[Y(0) \mid C = c] \\ &= E[Y(1) \mid C = c, Z = 1] - E[Y(0) \mid C = c, Z = 0] \\ &= E[Y \mid C = c, Z = 1] - E[Y \mid C = c, Z = 0] \end{aligned} \tag{4.3}$$

where the second and third equalities again follow from randomization and consistency, respectively. In other words, if adherence subgroup C were known, then the CACE could be computed as the ITT effect within the subset of the sample belonging to the Complier adherence subgroup. Our model for the data will therefore consist of two parts: a model for the unobserved adherence subgroup C and a model for the observed outcome Y within each combination of adherence subgroup C and treatment assignment Z. We first describe the data model and parameters for $C_i$, the adherence subgroup of the $i^{th}$ individual in the study. Since each individual can belong to only one adherence subgroup, Ci can be modeled by a Multinomial distribution of size 1, the form of which will depend on whether the strong access monotonicity assumption is made

$$C_i \sim Multinomial(\omega_c, \omega_n, \omega_a) \text{ (No strong access monotonicity)}$$
$$C_i \sim Multinomial(\omega_c, \omega_n) \text{ (strong access monotonicity)}$$

The parameter $\omega_t$ indicates the probability that an individual belongs to adherence subgroup $C = t$. While a two-parameter Multinomial model is more typically represented as a single-parameter Binomial model, we maintain the Multinomial representation for notational consistency between the two cases. Given this model for the adherence subgroup $C_i$, we must specify a prior distribution for its parameters. The Dirichlet distribution is a convenient prior for the parameters t of the Multinomial distribution

$$(\omega_c, \omega_n, \omega_a) \sim Dirichlet(\gamma_c, \gamma_n, \gamma_a) \text{ (No strong access mononicity)}$$
$$(\omega_c, \omega_n) \sim Dirichlet(\gamma_c, \gamma_n) \text{ (strong access monotonicity)}$$

The hyperparameters $\gamma_t$ control the shape of the prior distribution over the adher-

ence subgroup parameters. Thus the Multinomial-Dirichlet model for $C_i$ has either 3 parameters with 3 associated hyperparameters or 2 parameters with 2 associated hyperparameters. We now describe the models we will use for the outcome $Y$ conditional on $C$ and $Z$. We will focus on two commonly encountered types of outcome models: the binary outcome model and the normal outcome model. For the binary outcome model, the outcome for an individual belonging to adherence subgroup $C = t$ under treatment assignment $Z = z$ is modeled by the Bernoulli distribution

$$f(y \mid p_{tz}) \sim Bernoulli(p_{tz})$$

This distribution depends on a single parameter $p_{tz}$ representing the probability that the outcome will be equal to 1. Since it is a probability, this parameter must take values between 0 and 1. The Beta distribution can be used to flexibly model such quantities, and so our prior on $p_{tz}$ is given by

$$\pi(p_{tz}) \sim Beta(\alpha_{tz}, \beta_{tz})$$

where $\alpha_{tz}$ and $\beta_{tz}$ are the hyperparameters that determine the shape of the prior distribution. By specifying values for the hyperparameters, the analyst is able to change the shape of the prior, thereby giving greater prior belief to certain regions of the parameter space and lesser prior belief to other regions. For example, a $Beta(1, 1)$ prior would give equal prior belief to all possible parameter values, while a $Beta(12, 4)$ prior would place almost all of the prior belief on parameter values between 0.6 and 1. For the normal outcome model, the distribution of the outcome $Y$ for an individual belonging to adherence subgroup $C = t$ assigned to treatment $Z = z$ is modeled by the normal distribution, which depends on a mean parameter $\mu$ and a variance parameter $\sigma^2$

$$f(y \mid \mu_{tz}, \sigma^2_{tz}) \sim Normal(\mu_{tz}, \sigma^2_{tz})$$

Since there are two parameters determining the normal distribution, prior distributions must be specified for both parameters. A frequently used prior distribution is

the Normal-Inverse Gamma distribution

$$\pi(\sigma^2_{tz}) \sim Inv.Gamma(a_{tz}, b_{tz})$$
$$\pi(\mu_{tz} \mid \sigma^2_{tz}) \sim Normal(\theta_{tz}, \sigma^2_{tz}\tau_{tz})$$

As in the binary outcome case, the analyst can specify the hyperparameters to reflect prior information about the mean and variance parameters. For instance, there may be an existing body of scientific literature suggesting what a reasonable range for the mean of the outcome will be. The number of total parameters and hyperparameters related to the outcome model will vary depending on which assumptions are made in the analysis. In the case of the binary outcome model, if neither the exclusion restriction nor strong access monotonicity are assumed, then the full outcome model consists of 6 parameters, one for each combination of adherence type (Complier, Never Taker, and Always Taker) and treatment assignment (control or intervention). There are then 2 hyperparameters associated with each of the 6 parameters, giving 12 hyperparameters in total for the outcome model. If the exclusion restriction is assumed, then both the Never Takers assigned to the intervention group and the Never Takers assigned to the control group can be modeled with a single parameter, and similarly for the Always Takers. Thus the total number of outcome model parameters is reduced to 4, and the total number of outcome model hyperparameters reduced to 8. If the strong access monotonicity assumption is additionally made, then the parameters and hyperparameters for the Always Takers drop out of the model, leaving 3 parameters with 6 associated hyperparameters. Table 2 gives a complete listing of the model parameters and hyperparameters under the different combinations of outcome models and assumptions.

If each individuals adherence subgroup were known, then the model parameters could be easily estimated by performing separate analyses within each adherence subgroup. In particular, the CACE could be estimated by simply comparing the mean outcome among Compliers in the treatment group to the mean outcome of Compliers in the control group, as suggested by Equation 4.3. Since the adherence subgroups are not known in general, they must be imputed based on the data. Data augmentation [Tanner and Wong, 1987] is an algorithmic approach to performing Bayesian inference when missing data values must be imputed. Informally, the data

augmentation algorithm proceeds by alternating between two steps. In the first step, we fill in the missing adherence subgroup values based on our best guess as to their distribution given the observed data and current values for the model parameters. In the second step, we draw new values for the model parameters from the posterior distribution defined by our data model and assumed prior distributions, treating the imputed adherence subgroup values as if they were the (unobserved) true values. We then go back to the first step, using the new values of the model parameters to perform the next imputation of adherence subgroups. Going back and forth between these two steps creates a chain of parameter values that eventually converges to the desired posterior distribution. In this way, we obtain a sample of parameter values that can be used to approximate the posterior distribution of the model parameters given the observed data. Typically, several independent chains will be run, and a certain number of the initial iterations from each chain will be discarded to reflect the fact that the chain does not actually begin at the desired posterior distribution. The remaining samples from each chain will then be combined to form a single sample from the posterior distribution.

## *4.4 Implementation of the Data Augmentation Algorithm with the* `noncomplyR` *Package*

In this section we work through two applied examples using real datasets to show how the Bayesian adherence-based analysis described above can be implemented using the noncomplyR package in R, a free scientific programming language and statistical analysis environment [R Core Team, 2015]. Several software programs are available for running R. A basic version is available from the Comprehensive R Archive Network (https://cran.r-project.org), while a version with some additional user-friendly features is available from RStudio (https://www.rstudio.com). The `noncomplyR` package provides a set of functions for fitting non-adherence models for two commonly encountered types of outcomes: dichotomous outcomes and outcomes that are well-modeled by a Normal distribution. When using the `noncomplyR` package for the first time, it must first be installed. This can be accomplished by running the following code in the R console

```
1 install.packages("noncomplyR")
```

Once installed, the package can be loaded into an R session using the `library()` function

```
library("noncomplyR")
```

Running this command makes all of the functions in the `noncomplyR` package available for use. The main function in the `noncomplyR` package is the `compliance_chain` function, which performs a single run of the data augmentation algorithm for drawing from the posterior distribution of the model parameters. This is the most complex function in the `noncomplyR` package. We therefore spend some time now detailing each of the arguments that the user can supply to it. The `compliance_chain` function takes the following arguments: `dat`, `outcome_model`, `exclusion_restriction`, `strong_access`, `starting_values`, `hyper_parameters`, `n_iter`, and `n_burn`. The `dat` argument should be a data frame object containing three columns. The first column should contain the values for the outcome of interest. The second column should contain the treatment assignment variable, with assignment to the active intervention coded as 1 and assignment to control coded as 0. The third column should contain the variable indicating the treatment actually received, with 1 indicating that the intervention was received and 0 indicating that the intervention was not received. The `outcome_model` argument should be a character string, either binary or normal, indicating what outcome model should be used. The `exclusion_restriction` argument takes a logical value (either `TRUE` or `FALSE`) indicating whether the exclusion restriction assumption should be assumed or not when fitting the model. The `strong_access` argument also takes a logical value indicating whether the strong access monotonicity assumption should be made or not when fitting the model. The `starting_values` argument should be a vector of numbers giving the initial parameter values for starting the data augmentation algorithm. If no value is supplied, the initial parameter values are based on either a random draw from the prior distribution (for the binary outcome model) or the sample means and variances in the two treatment groups (for the normal outcome model). The `hyper_parameters` argument should be a vector of numbers that gives the hyper parameters for the prior distributions. If no value is supplied for this argument, the default is a non-informative uniform or reference prior. When supplying either the `starting_values` or `hyper_parameters`

arguments, the order of the values should follow the same conventions as those shown in Table 2. The `n_iter` argument is a number determining how many iterations of the data augmentation algorithm should be performed. If no value is supplied, then 10,000 iterations are performed as a default. The `n_burn` argument is a number determining how many of the initial iterations should be discarded. Discarding some portion of the initial iterations is standard practice, the rationale being that the early draws are less likely to be representative of the posterior distribution due to the time it takes for the algorithm to converge. The default is to discard the first 1,000 iterations. We now demonstrate the use of this function as well as several other useful functions by analyzing two data sets from actual RCTs. The first analysis will demonstrate the use of these functions with a binary outcome. The second analysis will demonstrate the use of these functions with a continuous outcome. The dataset we will use to demonstrate the analysis of binary outcome data is based on data described in Imbens and Rubin [1997]. The data come from a randomized trial performed in Indonesia investigating the effect of Vitamin A supplements on child mortality. The dataset is included in the `noncomplyR` package and is automatically saved into the R session as a data frame named `vitaminA` when the package is loaded. The dataset contains 23,682 observations on 3 dichotomous variables: `survived`, an indicator of whether the subject survived; `vitaminA_assigned`, an indicator of whether the subject was assigned to receive vitamin A supplements; and `vitaminA_received`, an indicator of whether the subject actually received vitamin A supplements. No individuals in the control group were recorded as having received the intervention. Thus, the strong access monotonicity assumption can be made. With only the Complier and Never Taker adherence subgroups, the data model is fully described by 6 probability parameters: $\omega_c$ and $\omega_n$, the probability of belonging to the Complier or Never Taker subgroups, respectively; $p_{c0}$ and $p_{c1}$, the probabilities of survival among Compliers assigned to either the control or intervention groups, respectively; and $p_{n0}$ and $p_{n1}$, the probabilities of survival among Never Takers assigned to either the control or intervention groups, respectively. We will perform two analyses, one with the exclusion restriction assumption and one without the exclusion restriction assumption. When the exclusion restriction assumption is made, the probabilities $p_{n0}$and $p_{n1}$ are assumed to be equal and the model therefore reduces to only 5 parameters. Both analyses will involve running a chain of the data augmentation algorithm for 10,000 iterations each.

The first 1000 iterations from each chain will be discarded, leaving 9,000 samples in each case to estimate the posterior distribution of the parameters. The two chains are fit using the following code

```
# Chain with Exclusion Restriction
chain1 <- compliance_chain(vitaminA, outcome_model = "binary", exclusion
    _restriction = T, strong_access = T)

# Chain without Exclusion Restriction
chain2 <- compliance_chain(vitaminA, outcome_model = "binary", exclusion
    _restriction = F, strong_access = T)
```

This code runs two data augmentation chains on the data and stores the results in matrix form in the two variables `chain1` and `chain2`. The columns of the matrices correspond to the different model parameters, ordered according to the conventions shown in Table 2, while each row represents a draw from the posterior distribution of the model parameters. Since the variables `n_iter`, `n_burn`, `starting_values`, and `hyper_parameters` were not specified, the `compliance_chain` function automatically supplies their default values. In this way, a Bayesian adherence-based analysis can be implemented with relatively simple syntax. Once a sample from the posterior distribution of all of the model parameters has been generated, the corresponding sample from the posterior of the CACE can be obtained by calling the `cace` function:

```
cace_1 <- cace(chain1, outcome_model = "binary", strong_access = T)

cace_2 <- cace(chain2, outcome_model = "binary", strong_access = T)
```

Inferential quantities of interest can now be directly calculated from these samples from the posterior distribution of the CACE. For convenience, the `noncomplyR` package provides the function `summarize_chain`, which automatically calculates several common inferential statistics. An example is shown below:

```
summarize_chain(cace_1)
# Posterior Mean: 0.003
# Posterior Median: 0.003
# Posterior 50% Credible Interval: (0, 0.005)
# Posterior 90% Credible Interval: (-0.001, 0.007)
# Posterior 95% Credible Interval: (-0.001, 0.007)
summarize_chain(cace_2)
# Posterior Mean: 0.003
# Posterior Median: 0.003
# Posterior 50% Credible Interval: (0.002, 0.004)
# Posterior 90% Credible Interval: (0.001, 0.005)
# Posterior 95% Credible Interval: (0.001, 0.005)
```

Histograms of the posterior distribution of the CACE from the two models are shown in Figure 4.1. The effect of the exclusion restriction assumption is evident from the difference in peakedness between the two histograms. The posterior histogram from the model with the exclusion restriction assumption has a steep, distinct peak. This implies that we can be reasonably certain that a point estimate such as the posterior mean or median will be a good summary of our beliefs about the magnitude and direction of the CACE given the observed data. In contrast, the posterior histogram from the model without the exclusion restriction assumption has a flatter, table-like appearance. This implies that we have roughly equal evidence given the observed data for all of the values of the CACE in the flat range of the posterior (in this case, from about -0.001 to about 0.006). This example illustrates two important aspects of a Bayesian adherence-based analysis. First, it shows the importance of the exclusion restriction assumption for making precise inference about the CACE. At the same time, it demonstrates how a Bayesian approach can still be used even in situations where the exclusion restriction cannot be assumed in order to learn from the data about the relative direction and magnitude of the CACE.

We now demonstrate how continuous outcome data from a randomized trial with non-adherence be analyzed using the `noncomplyR` package. The data in this section come from an RCT investigating the Early Start Denver Model (ESDM), an early intensive behavioral intervention for children with autism [Dawson et al., 2010]. While this data set is not available for public use, we describe the analysis and present the

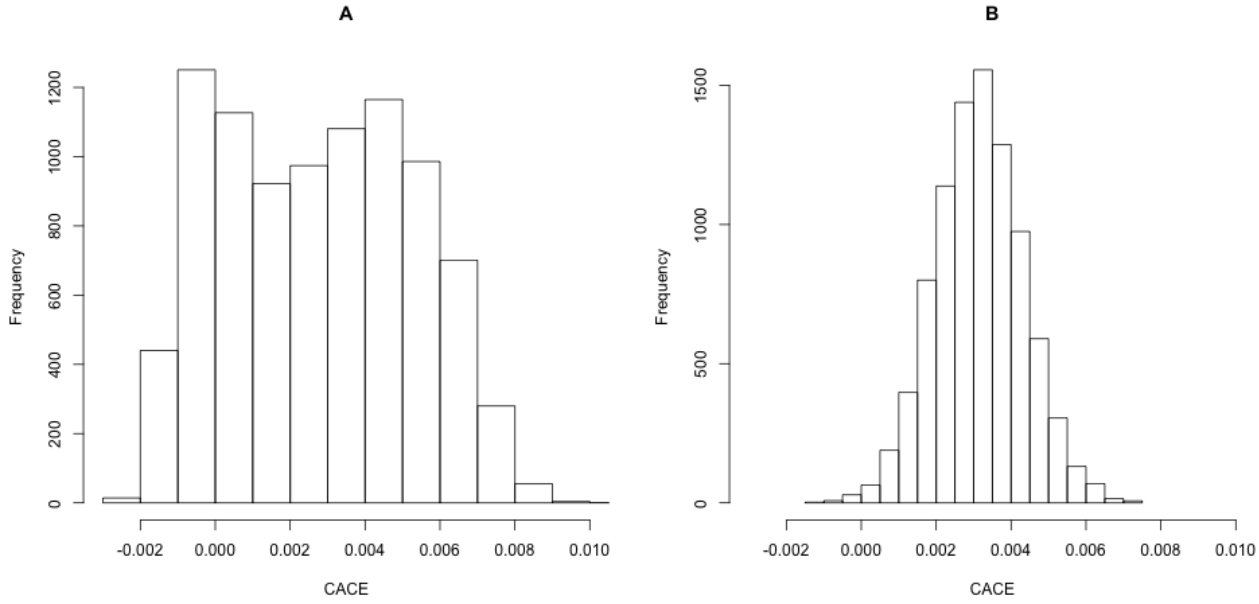



Figure 4.1: Posterior distributions of the Complier Average Causal Effect, both without the exclusion restriction assumption (left panel) and with the exclusion restriction assumption (right panel).

analysis code as an illustrative example. Participants in this study were randomized to one of two groups. The intervention group received one-on-one treatment according to the ESDM intervention protocol. The control group received information about community-based care opportunities. Information on several measures of interest was collected at baseline, one year post-randomization, and two years post-randomization. This analysis will focus on the effect of the ESDM intervention on the Mullen Scales of Early Learning (MSEL) composite score, a measure of cognitive functioning (Mullen, 1995). We will take as our outcome of interest the difference in the Mullen composite score at baseline and two years post-randomization

$$Y = \text{Mullen score at 2 year followup } - \text{ Mullen score at baseline}$$

Data were also collected on the number of weekly ESDM hours received by participants in the active intervention group. Based on these data, we can classify the observed adherence behavior of participants in the active intervention group. We created an observed adherence variable, $D$, for each individual in the active intervention group based on the cutoff value of 10 hours per week on average over the course of the study,

$$D = \begin{cases} 1 \text{ if ESDM hours } \geq 10 \\ 0 \text{ if ESDM hours } < 10 \end{cases}$$

In this study, the ESDM intervention was not available from community-based providers and thus inaccessible by individuals in the control group, so that D = 0 for all individuals assigned to the control group. Thus the strong access monotonicity assumption holds, and we only need to model the Complier and Never Taker adherence subgroups. In addition, we will make the exclusion restriction. Similar to the binary outcome example, we will run multiple chains of the data augmentation algorithm under different settings for 10,000 iterations each, discarding the first 1000 draws from each chain as burn-in. Unlike in the previous example, we will not use these multiple chains to demonstrate the effect of the exclusion restriction. Instead, we will demonstrate how the analyst can use the `noncomplyR` package to place informative priors on the model parameters, as well as the effect that these informative priors have

on the results of the analysis. We first run a single chain of the data augmentation algorithm using the default settings for many of the function arguments. The code for this chain is very similar to the code for the binary outcome example, but with the `outcome_model` argument set to "`normal`" instead of "`binary`":

```
chain_reference <- compliance_chain(esdm_dat, outcome_model = "normal",
    exclusion_restriction = T, strong_access = T)
```

Because no values were supplied to the `hyper_parameters` argument, the `compliance_chain` function automatically selects the reference prior for the Normal-Inverse Gamma distribution as the default prior. This prior can be thought of as an uninformative or flat prior that is suitable in situations where there is no strong background information on the problem at hand (for instance, in the early stages of investigation of a new type of intervention). In the case of behavioral interventions for children with autism, there is an existing body of literature on the efficacy of these interventions. Thus, researchers may wish to incorporate this prior information into their analysis. This can be accomplished through the selection of an informative prior. We now demonstrate how this can be done by supplying our own hyperparameter values to the `hyper_parameters` argument of the `compliance_chain` function. We note that these hyperparameter values have been chosen simply as an illustrative example and are not intended as a summary of the scientific consensus on the effects of early intervention for children with autism.

```
chain_informative <- compliance_chain(esdm_dat, outcome_model = "normal"
    , exclusion_restriction = T, strong_access = T,
hyper_parameters = c(1, 1, 5, 50, 5, 100, 15, 50, 5, 100, 2, 1, .1, 4) )
```

We note again that the order in which the hyperparameters are listed is important and will vary depending on the outcome model and the set of assumptions made. The conventions for the ordering of hyperparameters for both the binary and normal outcome models are detailed in Table 2. In this case, the chosen hyperparameters translate to a prior distribution for the CACE centered around 5 with a standard deviation of 10. This codifies a prior belief for a small beneficial effect of the ESDM intervention among the Complier subgroup, with sufficient uncertainty in the prior

to allow for a range of both negative and positive effects. This prior represents a balance between the desire to incorporate prior evidence for a positive effect of ESDM intervention and the need to let the data speak in the analysis. The difference in the resulting inference can be seen by examining posterior summaries from the two chains.

```
cace_reference <- cace(chain_reference, outcome_model = "normal", strong
    _access = T)

cace_informative <- cace(chain_informative, outcome_model = "normal",
    strong_access = T)

summarize_chain(cace_reference)
# Posterior Mean: 11.83704
# Posterior Median: 11.96331
# Posterior 50% Credible Interval: (8.245964, 15.49876)
# Posterior 90% Credible Interval: (2.820263, 20.66022)
# Posterior 95% Credible Interval: (0.7812756, 22.49723)

summarize_chain(cace_informative)
# Posterior Mean: 11.0748
# Posterior Median: 11.06049
# Posterior 50% Credible Interval: (8.051141, 14.08166)
# Posterior 90% Credible Interval: (3.577411, 18.58465)
# Posterior 95% Credible Interval: (2.010731, 20.07434)
```

While overall the inferences from the two posteriors are similar, there are some noticeable differences. The posterior mean of the CACE is slightly smaller under the informative prior compared to the reference prior, due to the weight placed by the informative prior on values of the CACE around the prior mean of 5. In addition, the informative prior results in posterior credible intervals that are tighter compared to the reference prior. This can be seen as well by graphically examining posterior densities of the two chains ( 4.2). Both densities have approximately the same posterior peak (somewhere between 11 and 12), but the posterior density based on the informative prior concentrates greater mass around this peak compared to the posterior based on the reference prior.

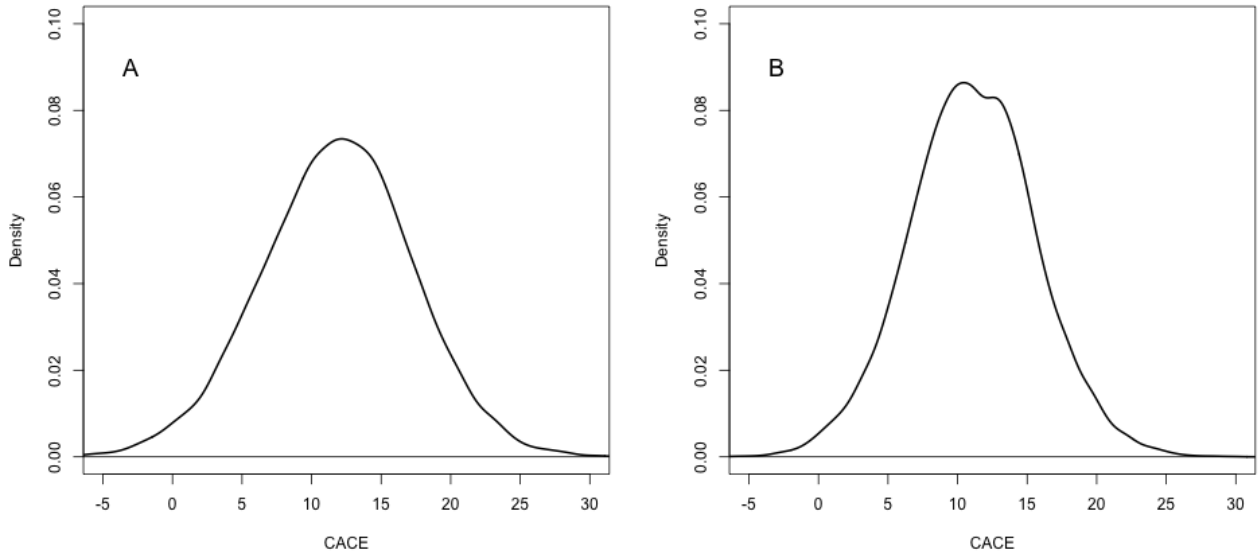



Figure 4.2: Left panel: Posterior distribution of the CACE from the ESDM dataset under an uninformative prior. Right panel: Posterior distribution of the CACE from the ESDM dataset under a user-specified informative prior.

These examples illustrate the general workflow of an adherence-based analysis using the `noncomplyR` package. This workflow can be summarized as follows
1. Organize the data set. The first column of the data set should contain the values of the outcome variable. The second column of the data set should contain the values for the treatment assignment. The third column should contain the values for the treatment actually received.
2. Use the `compliance_chain` function to generate a sample from the posterior distribution of the model parameters. The first argument to this function should be a data frame organized as described. The other arguments to the function can be determined by going through the following checklist:

a. Determine the appropriate outcome model

b. Determine if the exclusion restriction is reasonable to assume

c. Determine if the strong access monotonicity assumption is reasonable to assume

d. Decide if starting values other than the default should be used for the initial parameter values

e. Decide if an informative prior should be used

f. Decide on the total number of iterations of the data augmentation algorithm to perform

g. Decide how many of the initial iterations should be discarded

3. Use the `cace` function to transform the posterior sample from the full model parameters into a posterior sample for the CACE. 4. Use the `summarize_chain` function to obtain estimates of posterior quantities of interest.

### *4.5 Discussion*

Non-adherence to treatment assignment is an almost unavoidable issue in studies involving human participants. Because it cannot be accounted for in the study design stage and is beyond the control of the investigators once the study has begun,

non-adherence to treatment assignment is typically addressed at the analysis stage. However, taking information about non-adherence into account in the analysis must be done carefully due to the risk of introducing bias and confounding. With this tutorial, we have provided both a conceptual introduction to the causal analysis of data from randomized trials with non-adherence, and a practical guide to implementing these analyses with the `noncomplyR` package. As with any piece of statistical software, there is a trade-off between ease of use and flexibility. In developing this software, we have tended towards the former. The reason for doing this was so that researchers would be able to implement these methods by focusing on the higher-level aspects of the data and study design (e.g. Could individuals in the control group access the intervention? Is the outcome approximately Normally distributed?), rather than the mathematical details of the Bayesian sampling method. However, the ability to specify different sets of assumptions provides researchers control over the type of model being fit. In addition, by including the option to specify the hyperparameters that determine the prior distributions of the model parameters, we have given researchers flexibility with respect to the informativeness of the prior distributions, one of the key aspects of a Bayesian analysis.

# Chapter 5

# A SYSTEMATIC REVIEW OF EFFECTIVENESS STUDIES OF BEHAVIORAL INTERVENTIONS FOR CHILDREN WITH ASD

## *5.1 Introduction*

Autism Spectrum Disorder (ASD) has begun to be diagnosed at earlier ages [Daniels and Mandell, 2014]. A study of ten birth cohorts of California children found that the median age at autism diagnosis had decreased from approximately 4.4 years in 1992 to approximately 3.3 years in 2001 [Fountain et al., 2011]. This decrease has created a corresponding need for interventions designed for younger children with ASD. A seminal study of a comprehensive behavioral intervention conducted for 24 months at 40 hours per week for children under the age of four with ASD reported accelerated development and achievement of typical levels of cognitive functioning in a large proportion of the sample, spurring interest in intervention for individuals with ASD [Lovaas, 1987]. Since that time, evidence for the efficacy of a wide array of early behavioral interventions has been reported [Howlin and Magiati, 2009, Kasari and Patterson, 2012, Rogers and Vismara, 2008]. Comprehensive interventions for young children with ASD, defined as interventions that focused on treating the full range of core ASD symptoms rather than a single domain, have been demonstrated to have both short-term and long-term effects [Pickles et al., 2016, Estes et al., 2015, Howlin and Magiati, 2009]. Targeted interventions have also been found to improve outcomes on specific behaviors such as initiation of joint attention and peer response and may have short and longer term effects [Kasari and Patterson, 2012, Gulsrud et al., 2014]. Thus, research has established early intervention as efficacious for children with ASD, particularly when utilizing applied behavioral analysis principles [Rogers and Vismara, 2008, **?**]. Based on this evidence, practice guidelines have been developed and recommend that intervention begin as early as possible, further supporting the move toward earlier diagnosis and intervention [Zwaigenbaum et al., 2015]. Although there

is a consensus that behavioral interventions can improve outcomes among young children with ASD, extensive research has not yet been conducted to test the effectiveness of these interventions in community settings. The existing evidence largely relies on efficacy studies that estimate the effect of intervention conducted in highly-controlled environments under optimal circumstances. In these types of settings, children tend to receive intervention overseen by specialized, expert therapists at a high level of fidelity and consistency [Mandell et al., 2013]. In contrast, effectiveness studies estimate the effect of an intervention as delivered in real-world settings, outside of research clinics and laboratories from therapists who may have less specialized training [Gartlehner et al., 2006]. In community-based effectiveness studies intervention fidelity is less closely monitored and is likely to be lower than in efficacy studies. Children in efficacy studies may not be representative of the ASD population due to strict inclusion criteria and recruitment that takes place through academic research settings rather than typical community sources that employ minimal inclusion criteria [Kasari and Patterson, 2012]. The purpose of this systematic review and meta-analysis was to evaluate existing research on the effectiveness of early behavioral interventions administered in community settings, and where possible, combine evidence across studies through meta-analytic techniques. We found three outcome domains that were common across three or more studies and were included in a meta-analysis: adaptive behavior, cognitive functioning, and joint attention.

## *5.2 Methods*

Database Search The PubMed database was searched on July 30, 2016 for peer-reviewed manuscripts related to early interventions for children with ASD. The search terms used were (autism or autism spectrum disorder) AND effectiveness AND intervention. Study abstracts were reviewed to ensure studies met the inclusion criteria below. Studies were included if: 1. Participants were children with ASD below the average age of six. 2. The study used a group design. 3. The study focused on a behavioral intervention. 4. The intervention was delivered by a person (i.e., community service provider, parent, teachers). 5. The study was conducted in a community setting (i.e., a school, home, or community-based care center) 6. The study was published in English. Studies were excluded if: 1. The study used single-subject design. 2. The intervention utilized biomedical, complementary and alternative medicine, or

pharmaceutical agents. 3. Interventions were delivered through media (e.g. viewing a video). The search of the PubMed database yielded 345 papers for review. Of these, 263 papers were excluded from the review after an examination of the abstract revealed that inclusion criteria were not met. Bibliography searches of the papers identified by the PubMed database search and consultation with co-authors identified an additional three papers, bringing the total number of papers for inclusion to 23. Of these, thirteen papers were excluded after review of the manuscript revealed that the studies did not meet inclusion criteria, leaving ten total studies for review. Study Assessment and Abstracting Studies included in the review were assessed on study design, intervention protocol characteristics, outcomes measures, and estimated treatment effects. Two independent coders abstracted the results of these assessments on pre-written forms. The coding results were compared to ensure that agreement on at least 80% of items was achieved per study. If this 80% threshold was not reached, then the coders met to come to a consensus through discussion. Consultation with senior authors was conducted to resolve any further discrepancies. Effectiveness characteristics that were assessed include: 1. Inclusion/exclusion criteria: Participants were excluded only for conditions that would make delivery of the intervention unfeasible. 2. Setting: The primary setting for delivery of the intervention was a community setting.

Methodological characteristics that were assessed include:

1. Whether a control group was included.
2. Whether intervention assignment was randomized.
3. Whether fidelity to intervention was assessed.
4. The size of the sample.

Study participant characteristics that were assessed: 1. The ages of the participants.

2. The percentage of males and females.
3. The racial/ethnic characteristics of the participants.

Intervention characteristics that were assessed:

1. Whether the intervention was targeted or comprehensive.
2. Whether there was parental involvement in the delivery of the intervention.
3. The time period during which participants received intervention.

Outcome measures and estimated treatment effects were coded to compare study

findings and, where appropriate, combine to assess through meta-analysis. Outcome characteristics that were assessed:

1. Whether the outcome was measured as a continuous or categorical variable.
2. What domains the outcome assessed (i.e., cognitive functioning, core ASD symptoms, adaptive behavior).
3. How often the outcomes were measured during the study.

**Data Analysis**

Effect sizes were calculated for main outcome measures from each of the studies. We used the standardized mean difference as the effect size statistic for all studies [Lipsey and Wilson, 2001]. Although some studies reported study-specific effect sizes, we re-calculated effect size to put each on the same metric across studies and to provide standard error estimates so that confidence intervals could be calculated. As a result, the effect sizes reported in this review may differ from effect sizes reported in the original studies. Confidence intervals for the effect sizes were set at the 95% confidence level. Effect sizes for outcomes that assessed similar domains (adaptive behavior, cognitive development, and joint attention skills) were combined through meta-analysis techniques. Random effects meta-analysis models were used to account for variation in choice of outcome measure, intervention method, and demographic characteristics of study participants. Inverse variance weighting was used to account for varying sample sizes. Combined effect size estimates and 95% confidence intervals were calculated from the meta-analyses.

## *5.3 Results*

The full table of study characteristics appears in Appendix D. Characteristics related to the methodological quality of the studies are given in Table 5.1.

**Effectiveness Criteria**

Inclusion/exclusion criteria: Nine of the ten studies allowed for inclusion of children with comorbidity with other syndromes or disease unless that comorbidity would preclude delivery of the intervention as intended. Setting: Studies in this review either used a school (3), community-based care center (4), or the home (3) as the setting for implementation.

**Methodological Characteristics**

Control group: Seven of ten reviewed studies had a control condition for comparison

purposes [Kasari et al., 2014, Lawton and Kasari, 2012, Peters-Scheffer et al., 2010, Salt et al., 2002, Stadnick et al., 2015, Vivanti et al., 2014, Wetherby et al., 2014]. Three of ten studies used a single group pre- and post-test design [Eapen et al., 2013, Ingersoll and Wainer, 2013, Smith et al., 2010].

Randomization: Randomization of participants into active intervention or control groups was used in three of ten studies [Lawton and Kasari, 2012, Kasari et al., 2014, Wetherby et al., 2014]. Among the studies that had a control or comparison group but lacked randomization, reasons given for not using random assignment were either ethical concerns about withholding treatment or logistical concerns.

Fidelity: Seven of ten studies reported that clinician fidelity was assessed [Ingersoll and Wainer, 2013, Kasari and Patterson, 2012, Lawton and Kasari, 2012, Smith et al., 2010, Stadnick et al., 2015, Vivanti et al., 2014, Wetherby et al., 2014], either at the beginning of the study (three studies) or periodically throughout the intervention (four studies). However, only two studies reported results related to fidelity measures. Ingersoll and Wainer [2013] assessed parents ability to successfully implement the Project ImPACT intervention using the Project ImPACT Fidelity Rating Scale and reported pre- and post-intervention fidelity statistics for five subscales, as well as an overall score. Wetherby et al. [2014] rated fidelity at the beginning of the study and during approximately 20% of sessions.

Sample size: Sample sizes at the time of enrollment ranged from 16 to 112. The average sample size was 45.9 (SD=30.9). Seven of the ten studies utilized a control group. Sample sizes were evenly balanced between the intervention group and the comparison group in all seven controlled studies. The average sample size for the intervention groups was 27.5 (SD=18.0) compared to an average sample size of 25.8 (SD=16.1) for the comparison groups with a range from 5 to 30 for control groups and 9 to 51 for active intervention groups.

**Participant Characteristics**

Participant age: All studies reported participant age. The average age of participants ranged from 19.6 months to 55 months, with a median average age of 47.25 months.

Participant sex: Eight of ten studies reported the sex of the participants [Eapen et al., 2013, Ingersoll and Wainer, 2013, Kasari et al., 2014, Salt et al., 2002, Smith et al., 2010, Stadnick et al., 2015, Vivanti et al., 2014, Wetherby et al., 2014].

Participant race/ethnicity: Seven of ten studies reported the race/ethnicity of the

participants [Eapen et al., 2013, Ingersoll and Wainer, 2013, Kasari et al., 2014, Lawton and Kasari, 2012, Salt et al., 2002, Stadnick et al., 2015, Wetherby et al., 2014]. Studies did not report ethnicity and race in a consistent way. Some studies reported race/ethnicity as a binary characteristic (e.g. white vs. non-white) while other studies included a range of racial/ethnic categories.

**Intervention Characteristics**

Targeted vs Comprehensive: Six of ten studies were of comprehensive interventions [Eapen et al., 2013, Peters-Scheffer et al., 2010, Salt et al., 2002, Smith et al., 2010, Vivanti et al., 2014, Wetherby et al., 2014]. Of these, four were intensive interventions (over 15 hours per week) and two were low intensity, comprehensive interventions. Four of ten studies were of targeted interventions of low intensity (weekly sessions of 1-2 hours) [Ingersoll and Wainer, 2013, Kasari et al., 2014, Lawton and Kasari, 2012, Stadnick et al., 2015]. Parental involvement: Eight of ten studies included some parental component in the intervention [Ingersoll and Wainer, 2013, Kasari et al., 2014, Peters-Scheffer et al., 2010, Salt et al., 2002, Smith et al., 2010, Stadnick et al., 2015, Vivanti et al., 2014, Wetherby et al., 2014]. However, the extent to which primary caregivers were involved in implementing the intervention varied. In five studies, parents or caregivers were a primary intervention agent [Ingersoll and Wainer, 2013, Kasari et al., 2014, Smith et al., 2010, Stadnick et al., 2015, Wetherby et al., 2014]. In three studies, parents were offered training and encouraged to implement the intervention strategies in the home, but were not a primary intervention agent [Peters-Scheffer et al., 2010, Salt et al., 2002, Vivanti et al., 2014]. Intervention length: Length of intervention varied depending on the type of intervention under study. Comprehensive behavioral interventions had longer intervention durations, from 3 months to 12 months. Targeted interventions focused on teaching specific skills had shorter intervention durations, ranging from 5 to 12 weeks.

**Outcome Characteristics**

Types of outcomes: The types of outcomes assessed and the measures used varied across studies depending on the type of intervention under investigation. Comprehensive early interventions that aim to improve outcomes across a wide range of developmental characteristics assessed a wide range of developmental outcomes, including

intellectual functioning, social skills, severity of ASD symptoms, and development of adaptive behaviors. Targeted interventions tended to assess a more limited range of outcomes including the specific skill domains targeted by the intervention. The most frequently used outcome measure was the Vineland Adaptive Behavior Scales (VABS). Five of ten studies used the Vineland Adaptive Behavior Composite as an analysis outcome [Eapen et al., 2013, Peters-Scheffer et al., 2010, Salt et al., 2002, Vivanti et al., 2014, Wetherby et al., 2014], while one study used the Communication and Socialization subscales [Stadnick et al., 2015]. Six of ten studies measured cognitive development; three used the Mullen Scales of Early Learning [Eapen et al., 2013, Vivanti et al., 2014, Wetherby et al., 2014], two used measures of mental age [Peters-Scheffer et al., 2010, Smith et al., 2010], and one used the Bayley Scales of Infant Development [Salt et al., 2002]. Six of ten studies assessed ASD symptom severity as an outcome; two used the Autism Diagnostic Observation Schedule [Vivanti et al., 2014, Wetherby et al., 2014], one used the Social Communication Questionnaire [Eapen et al., 2013], two used the Social Responsiveness Scale [Ingersoll and Wainer, 2013, Smith et al., 2010], and one used the Pervasive Development Disorder in Mentally Retarded Persons scale [Peters-Scheffer et al., 2010]. Three of ten studies used measures of specific behaviors; all three included a measure of joint attention [Kasari et al., 2014, Lawton and Kasari, 2012, Salt et al., 2002].

### *5.4 Forest Plots and Meta-Analyses*

Adaptive behavior: For adaptive behavior as measured by the VABS, results from five studies were able to be included in the meta-analysis. One study that used VABS as an outcome measure was excluded due to lack of a control group for comparison. All five studies were controlled, but only one of the five was randomized. A forest plot of the effect size estimates and 95% confidence intervals, and the combined effect size estimate, is shown in Figure 5.1. Based on the estimated combined effect size of 0.67 (95% CI: (0.08, 1.26)), we conclude there is evidence for a moderate association between intervention and improvement in adaptive behavior.

Cognitive development: For cognitive development, results from three studies were able to be included in a meta-analysis. Two studies that used a measure of cognitive development as an outcome were excluded due to lack of a control group for

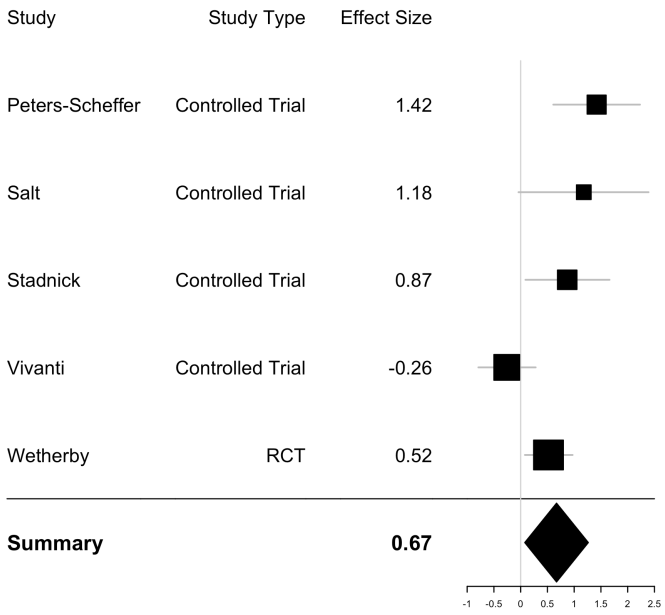



Figure 5.1: Forest plot of estimates for the Vineland Adaptive Behavior Scale outcome.

comparison, while another was excluded because data on cognitive development were only reported at baseline. All included studies were controlled, but only one was randomized. A forest plot of the effect size estimates and 95% confidence intervals, along with the combined effect size estimate, is shown in Figure 5.2. The estimated combined effect size was 0.8 (95% CI: (-0.01, 1.61)). We conclude that there is evidence for a moderate association between intervention and improvement in cognitive development, although this association was not significant at the 0.05 significance level.

Joint attention: For joint attention skills, results from three studies were able to be included in a meta-analysis. All of the studies were controlled, and two out of the three were randomized. A forest plot of the effect size estimates and 95% confidence intervals, along with the combined effect size estimate, are shown in Figure 5.3. Based on the estimated combined effect size of 1.06 (95% CI: (0.26, 1.86)), there is evidence to suggest a large association between behavioral therapy interventions for improving joint attention skills.

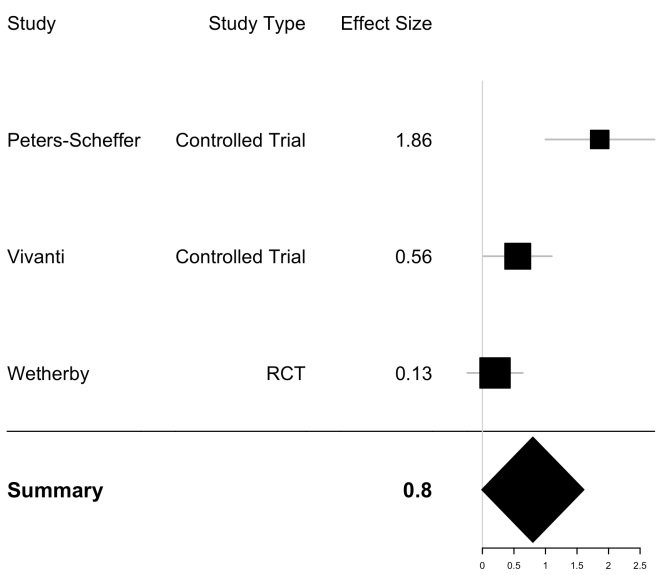



Figure 5.2: Forest plot of estimates for the cognitive development outcome.

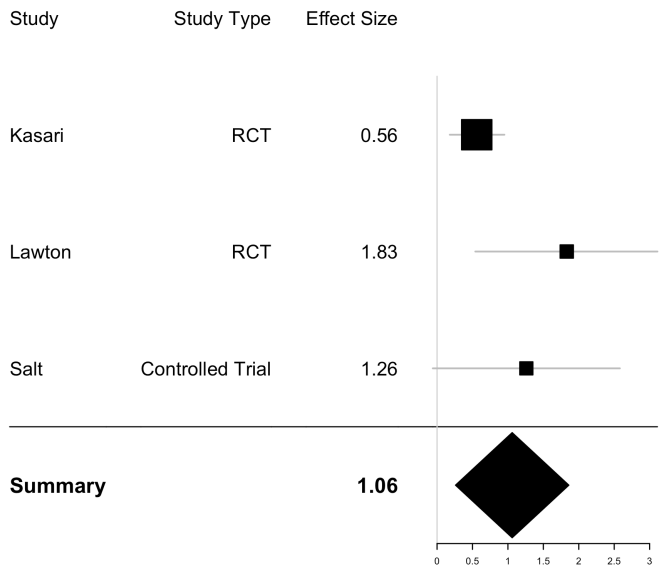



Figure 5.3: Forest plot of estimates for joint attention outcome.

Table 5.1: Characteristics of the review studies related to methodological quality.

| Author (Year) | Control Arm (Yes/No) | Randomization (Yes/No) | Sample Size | Fidelity Measured | Manualized Intervention |
|---|---|---|---|---|---|
| Eapen (2013) | No | No | 26 | No | Yes |
| Ingersoll (2013) | No | No | 27 | Yes | Yes |
| Kasari (2014) | Yes | Yes | 112 | Yes | Yes |
| Lawton (2012) | Yes | Yes | 16 | Yes | Yes |
| Peters-Scheffer (2010) | Yes | No | 34 | No | Yes |
| Salt (2002) | Yes | No | 20 | No | No |
| Smith (2010) | No | No | 53 | Yes | Yes |
| Stadnick (2015) | Yes | No | 30 | Yes | Yes |
| Vivanti (2014) | Yes | No | 59 | Yes | Yes |
| Wetherby (2014) | Yes | Yes | 82 | Yes | Yes |

## *5.5 Discussion*

Evidence from studies of early ASD interventions conducted in community settings provide strong meta-analytic evidence that behavioral interventions for young children with ASD are associated with improved adaptive functioning and joint attention. Weaker evidence was found for improved cognitive function. These results are consistent with those of highly controlled efficacy trials that also demonstrate evidence of significantly improve developmental outcomes for children with ASD who participate in early ASD intervention. Previous reviews have called for greater methodological rigor in studies of early ASD intervention (Howlin  Magiati, 2009) (Warren et al., 2011). Although many of the studies conducted in community-based settings had control groups for comparison, only three of ten featured randomization of participants to treatment groups. Lack of randomization may be due to reluctance on the part of community care providers to exclude a subset of individuals from the intervention program being tested. This sometimes results in the perception that community care provider goals are in conflict with the goals of academic researchers (Vivanti et al., 2014). Ethical concerns over withholding treatment were often cited as a primary reason for using a wait-list control group rather than utilizing a randomized control group. As researchers move increasingly toward implementing studies in community-based settings, attention to establishing collaborative partnerships and

communication about the critical role of randomization for determining effectiveness may be able to increase the number of randomized trials in community-based studies.

However, even conclusions from methodologically strong studies will be limited if the sample size is small. Issues related to the difficulty of conducting large-scale studies of ASD intervention are well-known [Howlin and Magiati, 2009, Wetherby et al., 2014]. Sample sizes in studies of behavioral interventions for children with ASD are constrained by both the nature of these interventions and the cost of conducting these types of studies. The interventions require a large time commitment from those implementing and participating, thus restricting the number of children that can feasibly be enrolled. In addition, finding a sufficient number of children with ASD to enroll in a small geographic area can be difficult due to the prevalence of ASD in the overall population. Given these natural limitations on study size, combining quantitative evidence across studies through meta-analyses will be crucial for understanding the impact of these interventions. Meta-analytic techniques provide a way to improve the strength of the evidence for behavioral interventions in the near-term future, without the need for drastic increases in the sample sizes of individual studies. However, the key consideration in performing a meta-analysis is whether it is appropriate to combine effect size estimates across studies. Future investigators should consider using a core battery of common outcome measures and include crucial information when report study results. Specifically, the size of the sample at recruitment as well as the number of individuals included in each analysis should be explicitly stated. Means and standard deviations of outcome measures at baseline and all follow-up times should be reported. Since treatment effects are often quantified in terms of the difference between the outcome measured at follow-up and the outcome measured at baseline, means and standard deviations of these differences are also needed. Planning, conducting, and disseminating results with an eye toward inclusion in a systematic review or meta-analysis will support progress in this critical area. One concern when moving from efficacy to effectiveness studies is whether fidelity to the intervention is maintained. Implementation fidelity may need special attention in effectiveness studies due to the challenges of conducting research in community settings. Progress in this area will rely on creative solutions for measuring implementation fidelity and for studies that include fidelity measures, reporting this data will improve future effectiveness studies. The majority of studies reported some form of fidelity data but this often

took the form of a single summary measure of fidelity assessed once during the study. Less common were detailed statistics on the nuanced components of fidelity such as adherence, dose, and quality of intervention delivery (Proctor et al., 2011) as well as longitudinal fidelity data collection throughout the duration of the study. As behavioral intervention studies move from tightly controlled environments to community settings, detailed information on the extent to which the interventions as developed were actually delivered will aid in understanding their effectiveness. Additionally, this information could contribute to the identification of ASD sub-types as well as the refinement of existing interventions [Stahmer et al., 2016]. However, as discussed in Kendall and Beidas [2007], researchers may need to allow for greater flexibility in fidelity (p. 13) when implementing interventions in community settings.

### *5.6 Limitations*

The meta-analyses presented in this review provide evidence for the effectiveness of early behavioral interventions. However, the small number of studies that were able to be included in the meta-analyses was a limitation. As a result, the combined effect size estimates are not as precise as they would be with a greater number of studies. As more controlled trials of behavioral interventions in real-world settings accumulate in the literature, it will be important to perform updated meta-analyses to obtain more precise effect-size estimates. The studies used in the meta-analyses were all controlled trials. However, most studies were not randomized. Future reviews with more randomized trials could restrict meta-analyses to only those studies with randomization, thereby increasing the reliability of the results.

### *5.7 Conclusion*

Preliminary evidence suggests that early ASD interventions may be effective in improving adaptive functioning and joint attention, and possibly IQ. Our review demonstrated a need for more community-based trials of early ASD intervention and for continued attention to increased methodological rigor in future effectiveness studies, with an emphasis on utilizing random assignment when possible and increasing sample sizes. Randomization may be challenging to implement in community settings due to concerns about withholding potentially effective treatment from children but

as Vivanti et al. [2014] stressed, establishing partnerships and shared understanding between academic researchers and community care providers may increase opportunities for random assignment. As more studies focused on effectiveness are conducted, meta-analytic techniques can be used to address issues related to small sample sizes and to increase confidence in the overall evidence.

# Chapter 6

# DISCUSSION

In Chapter 1, I considered the case of a single-site RCT with multiple versions of control and potential treatment-arm noncompliance. The framework for modeling the versions of control was based on a particular pattern of control group heterogeneity often seen in trials of behavioral interventions. This pattern of control group heterogeneity motivated the extensions and modifications of the consistency and monotonicity assumptions beyond their usual formualations. I demonstrated that the causal estimands $\Delta_{01}$ and $\Delta_{21}$ remain unidentified even with these modified assumptions and the exclusion restriction assumption. I then derived identifiable bounds on these quantities, and showed how estimation and inference for these bounds can be performed. An area for future research would be seeing how baseline covariates can possibly tighten the bounds. Another area for future research is extending the results obtained here for a binary outcome to the case of a continuous but bounded outcome.

In Chapter 2, I considered identification of these causal estimands when data are available from multiple populations, using the case of a multi-site RCT as a motivating example. I showed that point identification of these causal estimands is possible with data from multiple populations when a constant treatment efffect is assumed across the populations so long as the exclusion restriction holds in each site and the distributions of the principal strata across sites satisfy certain conditions. I also showed how tighter bounds on $\Delta_{01}$ and $\Delta_{21}$ can be obtained using data from multiple populations when the exclusion restriction assumption does not hold. Areas for future research include evaluation of the theoretical properties of these tighter bounds, as well as extension of this framework to the case where outcomes within the same site are considered correlated.

In Chapter 3, I showed that previously proposed likelihood-based methods for adjusting for partial compliance to treatment in RCTs lead to inconsistent estimation of their target of inference under two of the most commonly specified outcome mod-

els, the linear regression model for a normally distributed outcome and the logistic regression model for a binary outcome. I proposed an alternative likelihood-based method, but demonstrated that it has issues with consistency and bias as well. The conclusion I reached was that the problems lie not so much with the estimation procedures but more so with the partial compliance models themselves. These models simply rely too much on the unidentified joint distribution of potential compliance behaviors. As such, they are more likely than not to produce misleading results and should be avoided in practice.

In Chapters 4 and 5, I presented more applied work aimed at two important but sometimes overlooked roles for statisticians when working with collaborators: educator and fellow scientist. In Chapter 4, I presented a tutorial along with statistical software for impementing a Bayesian approach to causal inference in RCTs with binary non-compliance. Both the tutorial and the software were written with the intended audience of a researcher or scientist interested in these methods but without the statistical or mathematical training necessary to fully understand the source materials or implement the methods themselves. Such researchers likely make up a significant portion of the people actually performing data analysis in the real world. Hence, I believe it is vital that all statisticians, not just those in charge of classrooms, embrace the role of educator. At the same time, I believe it is important that statisticians make an effort to understand the science behind the projects that they are involved in. Such efforts will not only lead to improved analyses, but will also help prevent statisticians from being viewed by their collaborators as mere "number crunchers".

# BIBLIOGRAPHY

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439): 1171–1176, 1997.

Francesco Bartolucci and Leonardo Grilli. Modeling partial compliance through copulas in a principal stratification framework. *Journal of the American Statistical Association*, 106(494):469–479, 2011.

P. Bibby, S. Eikeseth, N. Martin, O. Mudford, and D. Reeves. Progress and outcomes for children with autism receiving parent-managed intensive interventions. *Research in Developmental Disabilities.*, 22:425–447, 2001.

Blai Bonet. Instrumentality tests revisited. *ArXiv e-prints*, January 2013.

Jing Cheng and Dylan S. Small. Bounds on causal effects in three-arm trials with noncompliance. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(5):815–836, 2006.

SR Cole and CE Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20:3–5, 2009.

AM Daniels and DS Mandell. Explaining differences in age at autism spectrum disorder diagnosis: a critical review. *Autism*, 18(5):583–597, 2014.

Geraldine Dawson, Sally J Rogers, Jeff Munson, Milani Smith, Jamie Winter, Jessica Greenson, Amy Donaldson, and Jennifer Varley. Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. *Pediatrics*, 125(1):e17–23, 2010.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

V. Eapen, R. Crncec, and A. Walter. Clinical outcomes of an early intervention program for preschool children with autism spectrum disorder in a community group setting. *BMC Pediatrics*, 13(3), 2013.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

Bradley Efron and D. Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, March 1991.

A. Estes, J. Munson, SJ. Rogers, J. Greenson, J. Winter, and G. Dawson. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child  Adolescent Psychiatry*, 54(7): 580–587, 2015.

Yanqin Fan and Andrew Patton. Copulas in econometrics. *Annual Review of Economics*, 6:179–200, 2014.

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification and confidence sets for functionals of the joint distribution of potential outcomes. *Working Paper*, 2014.

JL Fleiss. Analysis of data from multiclinic trials. *Controlled Clinical Trials*, 7(4): 267–275, 1986.

C Fountain, MD King, and PS Bearman. Age of diagnosis for autism: individual and community factors across 10 birth cohorts. *Journal of Epidemiology and Community Health*, 65(6):503–510, 2011.

Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002a.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, March 2002b.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of Clinical Trials*. Springer, Switzerland, 2015.

G. Gartlehner, RA. Hansen, D. Nissman, KN. Lohr, and TS. Carey. Criteria for distinguishing effectiveness from efficacy trials in systematic reviews. 2006.

AC. Gulsrud, C. Kasari, S. Freeman, and T. Paparella. Children with autism's response to novel stimuli while participating in interventions targeting joint attention or symbolic play skills. *Autism*, 11(6):535–546, 2007.

AC. Gulsrud, GS. Hellemann, SFN. Freeman, and C. Kasari. Two to ten years: developmental trajectories of joint attention in children with asd who received targeted social communication interventions. *Autism Research*, 7:207–215, 2014.

Paul Gustafson. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in Medicine*, 24:1203–1217, 2005a.

Paul Gustafson. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2), 2005b.

Paul Gustafson. *Bayesian inference for partially identified models: Exploring the limits of limited data*. Monographs on statistics and applied probability. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2015.

Raiden B. Hasegawa, Sameer K. Deshpande, Dylan S. Small, and Paul R. Rosenbaum. "causal inference with two versions of treatment". *ArXiv e-prints*, May 2017.

Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiology*, 22(3):368–377, 2011.

Keisuke Hirano, Guido W Imbens, Donald B Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1): 69–88, 2000.

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

Joel L Horowitz and Charles F Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84, 2000.

P. Howlin and I. Magiati. Review of early intensive behavioral interventions for children with autism. *AJIDD*, 114(1):23–41, 2009.

MG Hudgens and ME Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

S L Hui and S D Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171, March 1980.

Joseph G. Ibrahim and Sanford Weisberg. Incomplete data in generalized linear models with continuous covariates. *Australian and New Zealand Journal of statistics*, 34(3):461–470, 1992.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327, 1997.

BR. Ingersoll and A. Dvortcsak. *Teaching social communication to children with autism: a practitioner's guide to parent training.* 2010. The Guilford Press, New York, 2010.

BR. Ingersoll and AL. Wainer. Pilot study of a school-based parent training program for preschoolers with asd. *Autism*, 17(4):434–448, 2013.

Piotr Jaworski, Fabrizio Durante, Wolfgang Haerdle, and Tomasz Rychlik, editors. *Copula Theory and its Applications*, 2010. Springer-Verlag.

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

Hui Jin and Donald B. Rubin. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103 (481):101–111, March 2008.

Geoffrey Jones, Wesley O. Johnson, Timothy E. Hanson, and Ronald Christensen. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66:855–863, 2010.

Peter Juni, Douglas G Altman, and Matthias Egger. Systematic reviews in health care: assesing the quality of controlled clinical trials. *British Medical Journal*, 323 (7303):42–46, 2001.

C. Kasari and S. Patterson. Interventions addressing social impairment in autism. *Current Psychiatry Reports*, 14(6):713–725, 2012.

C. Kasari, K. Lawton, W. Shih, TV. Barker, R. Landa, C. Lord, F. Orlich, B. King, AM. Wetherby, and D. Senturk. Caregiver-mediated intervention for low-resourced pre-schoolers with autism: an rct. *Pediatrics*, 134(1):e72–e79, 2014.

PC. Kendall and RS. Beidas. Smoothing the trail for dissemination of evidence-based practices for youth: flexibility within fidelity. *Professional Psychology: Research and Practice*, 38(1):13–20, 2007.

Roger Koenker. *Quantile Regression*. Cambridge University Press, New York, NY, USA, 2005.

Helena C Kraemer and Thomas N Robinson. Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? *Contemporary Clinical Trials*, 26(5):518–529, 2005.

K. Lawton and C. Kasari. Teacher-implemented joint attention intervention: pilot randomized controlled study for preschoolers with autism. *JCCP*, 80(4):687–693, 2012.

MW Lipsey and DB Wilson. *Practical Meta-Analysis*. Applied Social Research Methods Series. Sage Publications, Thousand Oaks, CA, 2001.

A Localio, JA Berlin, TR Ten Have, and SE Kimmel. Adjustments for center in multicenter studies: An overview. *Annals of Internal Medicine*, 135 (2):112–123, 2001. doi: 10.7326/0003-4819-135-2-200107170-00012. URL + `http://dx.doi.org/10.7326/0003-4819-135-2-200107170-00012`.

Dustin M Long and Michael G Hudgens. Sharpening bounds on principal effects with covariates. *Biometrics*, 69:812–819, 2013.

Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2): 226–233, 1982.

OI Lovaas. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *JCCP*, 55(1):3–9, 1987.

OI Lovaas. *Teaching individuals with developmental delays: basic intervention techniques*. Pro-Ed, Austin, Texas, 2003.

Yan Ma, Jason Roy, and Bess Marcus. Causal models for randomized trials with two active treatments and continuous compliance. *Statistics in Medicine*, 30:2349–2362, May 2011.

DS Mandell, AC Stahmer, Sujie Shin, Ming Xie, E Reisinger, and SC Marcus. The role of treatment fidelity on outcomes during a randomized field trial of an autism intervention. *Autism*, 17(3):281–295, 2013.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Joesph Millum and Christine Grady. The ethics of placebo-controlled trials: methodological justifications. *Contemporary Clinical Trials*, 36:510–514, 2013.

Jiri Moutesek and Bernd Gartner. *Understanding and Using Linear Programming*. Springer, Berlin, 2005.

E.M. Mullen. *Mullen Scales of Early Learning: AGS Edition*. American Guidance Service, Circle Pines, MN, 1995.

Kevin M Murphy and Robert H Topel. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 20(1):88–97, 2002.

JC Nash. The (dantzig) simplex method for linear programming. *Computing in Science and Engineering*, 2(1):29–31, 2000.

National Autism Center. Findings and conclusions: national standards project, phase 2. Report, 2015.

JA Nelder and RWM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

J Neyman. On the application of probability theory to agricultural experiments. *Statistical Sciences*, 5(4):465–472, 1923 (1990).

S. Ozonoff and K. Cathcart. Effectiveness of a home program intervention for young children with autism. *JADD*, 28(1):25–32, 1998.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, New York, NY, USA, 2000.

Judea Pearl. On the consistency rule in causal inference: axiom, definition, assumption or theorem? *Epidemiology*, 21(6):872–875, 2010.

N. Peters-Scheffer, R. Didden, M. Mulders, and H. Korzilius. Low intensity behavioral treatment supplementing preschool services for young children with autism spectrum disorders and severe to mild intellectual disability. *Research in Developmental Disabilities.*, 31(6):1678–1684, 2010.

Maya L Petersen. Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology*, 22(3):378–381, 2011.

A. Pickles, A. Le Couteur, K. Leadbitter, E. Salomone, R. Cole-Fletcher, H. Tobin, I. Gammer, J. Lowry, G. Vamvakas, S. Byford, C. Aldred, V. Slonims, H. McConachie, P. Howlin, JR. Parr, T. Charman, and J. Green. Parent-mediated social communication therapy for young children with autism (pact): long-term follow-up of a randomised controlled trial. *The Lancet*, 2016.

R Core Team. `http://fs.fish.govt.nz/Page.aspx?pk=7sc=SUR`, 2015.

B. Reichow, F. Volkmar, and D. Cicchetti. Development of the evaluative method for evaluating and determining evidence-based practices in autism. *JADD*, 38: 1311–1319, 2008.

Thomas S Richardson, Robin J Evans, and James M Robins. Transparent parametrizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

S. Rogers and L. Vismara. Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child  Adolescent Psychology*, 37(1):8–38, 2008.

Sally J Rogers, Annette Estes, Catherine Lord, Laurie Vismara, Jamie Winter, Annette Fitzpatrick, Mengye Guo, and Geraldine Dawson. Effects of a brief early start denver model (esdm)-based parent intervention on toddlers at risk for autism spectrum disorders:a randomized controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51(10):1052–1065, 2012.

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

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371): 591–593, 1980.

Jose M Sallan, Oriol Lordan, and Vicenc Fernandez. OmniaScience, 2015.

J. Salt, J. Shemilt, V. Sellars, S. Boyd, T. Coulson, and S. McCool. The scottish centre for autism preschool treatment programme: Ii: the results of a controlled treatment outcome study. *Autism*, 6(1):33–46, 2002.

Sharon Schwartz, Nicolle M. Gatto, and Ulka B. Campbell. Extending the sufficient component cause model to describe the stable unit treatment value assumption (sutva). *Epidemiologic Perspectives and Innovations*, 9(3), 2012.

IM. Smith, LK. Koegel, RL. Koegel, DA. Openden, KL. Fossum, and SE. Bryson. Effectiveness of a novel community-based early intervention model for children with autism spectrum disorder. *AJIDD*, 115(6):504–523, 2010.

S. Sparrow, D. Balla, and D. Cicchetti. *Vineland adaptive behavior scales: second edition.* American Guidance Service, Shoreview, MN, 2005.

NA. Stadnick, A. Stahmer, and L. Brookman-Frazee. Preliminary effectiveness of project impact: a parent-mediated intervention for children with autism spectrum disorder delivered in a community program. *JADD*, 45:2092–2104, 2015.

A. Stahmer, J. Suhrheinrich, and DS Mandell. The importance of characterizing intervention for individuals with autism. *Autism*, 20(4):386–387, 2016.

Martin A. Tanner and WH Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

Leslie Taylor and Xiao-Hua Zhou. Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics*, 65(1):88–95, 2009.

Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):3–5, 2009.

Tyler J VanderWeele and Miguel A Hernan. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.

G. Vivanti, J. Paynter, E. Duncan, H. Fothergill, C. Dissanayake, SJ. Rogers, and the Victorian ASELCC Team. Effectiveness and feasibility of the early start denver model implemented in a group-based community childcare setting. *JADD*, 44: 3140–3153, 2014.

Z. Warren, ML. McPheeters, N. Sathe, JH. Foss-Feig, A. Glasser, and J. Veenstra-VanderWeele. A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics*, 127(E1303), 2011.

AM. Wetherby, W. Guthrie, J. Woods, C. Schatschneider, RD. Holland, L. Morgan, and C. Lord. Parent-implemented social intervention for toddlers with autism: an rct. *Pediatrics*, 134:1084–1093, 2014.

C Wong, SL Odom, KA Hume, AW Cox, A Fettig, S Kucharczyk, ME Brock, JB Plavnick, VP Fleury, and TR Schultz. Evidence-based practices for children, youth, and young adults with autism spectrum disorder: a comprehensive review. *JADD*, 45:1951–1966, 2015.

Fan Yang and Dylan S Small. Using post-outcome measurement information in censoring-by-death problems. *Journal of the Royal Statistical Society, Series B*, 78(1):299–318, 2016.

Linda HY Yau and Roderick J Little. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, 94(456):1232–1244, 2001.

Xiao-Hua Zhou and Sierra M Li. Itt analysis of randomized encouragement design studies with missing data. *Statistics in Medicine*, 25:2737–2761, 2006.

L. Zwaigenbaum, ML. Bauman, R. Choueiri, C. Kasari, A. Carter, D. Granpeesheh, Z. Mailoux, S. Smith Roley, S. Wagner, D. Fein, K. Pierce, T. Buie, PA. Davis, C. Newschaffer, D. Robins, AM. Wetherby, WL. Stone, N. Yirmiya, A. Estes, RA. Hansen, JC. McPartland, and MR. Natowicz. Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics*, 136(Supplement 1):S60–S81, 2015.

# Appendix A

## *A.1 Proofs*

*Proof of Lemma 1*

*Consistency.*

Each estimator $\hat{q}_{dy\cdot z}$ can be written as a ratio of sample means

$$\begin{aligned}\hat{q}_{dy\cdot z} &= \hat{P}(Y=y, D=d \mid Z=z) \\ &= \frac{\sum_{i=1}^{n} \mathbb{1}(Y_i=y)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)}{\sum_{i=1}^{n} \mathbb{1}(Z_i=z)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i=y)\mathbb{1}(D_i=d)\mathbb{1}(Z_i=z)}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i=z)}.\end{aligned}$$

By the Weak Law of Large Numbers (WLLN), the numerator converges in probability to its expectation,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D_i=d)\mathbb{1}(Y_i=y)\mathbb{1}(Z_i=z) &\xrightarrow{p} E[\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y)\mathbb{1}(Z_i=z)] \\ &= E_Z[\mathbb{1}(Z_i=z)E[\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y) \mid Z_i=z]] \\ &= P(Z_i=z)P(Y_i=y, D_i=d \mid Z_i=z) \\ &= P(Z_i=z)q_{dy\cdot z}.\end{aligned}$$

Similarly, by the WLLN the denominator converges in probability to its expectation

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i=z) &\xrightarrow{p} E[\mathbb{1}(Z_i=z)] \\ &= P(Z_i=z).\end{aligned}$$

Thus by Slutksy's Theorem and the Continuous Mapping Theorem, the ratio of these quantities converges in probability to the ratio of their respective limits,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y)\mathbb{1}(Z_i=z)}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z)} \xrightarrow{p} (P(Z_i=z))^{-1}P(Z_i=z)P(Y_i=y,D_i=d \mid Z_i=z)$$
$$= q_{dy\cdot z}.$$

*Asymptotic Normality.*
Let $\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=1)$ and $\hat{p}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=0)$. The quantities $\hat{q}_{dy\cdot z} - q_{dy\cdot z}$ can be written

$$\begin{aligned}
\hat{q}_{dy\cdot z} - q_{dy\cdot z} &= \frac{\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)1(Z_i=z)1(Y_i=y)}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z)} - q_{dy\cdot z} \\
&= (\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z))^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y)\mathbb{1}(Z_i=z) - q_{dy\cdot z} \\
&= (\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z))^{-1}(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y)\mathbb{1}(Z_i=z) - q_{dy\cdot z}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z)) \\
&= (\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i=z))^{-1}\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(D_i=d)\mathbb{1}(Y_i=y) - q_{dy\cdot z})\mathbb{1}(Z_i=z) \\
&\equiv (\hat{p}_z)^{-1}\tilde{q}_{dy\cdot z}.
\end{aligned}$$

and thus the vector $\hat{\boldsymbol{q}} - \boldsymbol{q}$ can be written as

$$\hat{\boldsymbol{q}} - \boldsymbol{q} = \hat{\boldsymbol{P}}^{-1}\tilde{\boldsymbol{q}}$$

where $\hat{\boldsymbol{P}}$ is the $8 \times 8$ diagonal matrix with entries

$$\hat{\boldsymbol{P}} = \begin{pmatrix} \hat{p}_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{p}_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{p}_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{p}_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{p}_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{p}_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{p}_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{p}_1 \end{pmatrix}$$

. The elements of $\tilde{\boldsymbol{q}}$ are sample means of the centered random variables $X_{dy\cdot z} = (\mathbb{1}(D = d)\mathbb{1}(Y = y) - q_{dy\cdot z})\mathbb{1}(Z = z)$ which have expectation and variance

$$E[X_{dy\cdot z}] = 0$$

$$Var[X_{dy\cdot z}] = p_z q_{dy\cdot z}(1 - q_{dy\cdot z}).$$

For $d, y, d', y'$ where either $d \neq d'$, $y \neq y'$ or both, the covariance is given by

$$Cov(X_{dy\cdot z}, X_{d'y'\cdot z}) = -p_z q_{dy\cdot z} q_{d'y'\cdot z}.$$

By the CLT,

$$\sqrt{n}\tilde{\boldsymbol{q}} \xrightarrow{d} N_8(\mathbf{0}, \Sigma_q^*)$$

where

$$\Sigma_q^* = \begin{pmatrix} p_0\Sigma_0 & \mathbf{0} \\ \mathbf{0} & p_1\Sigma_1 \end{pmatrix}$$

By the WLLN

$$\hat{p}_z \xrightarrow{p} p_z$$

and thus by the Continuous Mapping Theorem

$$\hat{\boldsymbol{P}}^{-1} \xrightarrow{p} \boldsymbol{P}^{-1}$$

where

$$\boldsymbol{P} = \begin{pmatrix} p_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_1 \end{pmatrix}.$$

We can therefore conclude that

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q}) &= \hat{\boldsymbol{P}}^{-1}\sqrt{n}\tilde{\boldsymbol{q}} \\ &\xrightarrow{d} N_8(\boldsymbol{0}, \boldsymbol{P}^{-1}\Sigma_q^*\boldsymbol{P}^{-1}) \\ &= N_8(\boldsymbol{0}, \Sigma_q) \end{aligned}$$

by Slutsky's theorem. ■

*Proof of Lemma 2*

*Identifiability of Causal Parameters $\omega_{d_0 d_1}$ and $p_{d_0 d_1 \cdot z}$*

First, we note that the quantities

$$q_{dy\cdot z} = P(Y = y, D = d \mid Z = z)$$

are defined completely in terms of the observed triples $(Y, D, Z)$ and thus are self-evidently identified. We now show that the causal parameters $\boldsymbol{\omega} = (\omega_{00}, \omega_{01}, \omega_{11}, \omega_{21})$ which determine the distribution of the principal strata are point identified. For $d = 1, 2$,

$$
\begin{aligned}
\omega_{d1} &= P(D(0) = d, D(1) = 1) \\
&= P(D(0) = d)P(D(1) = 1 \mid D(0) = d) \\
&= P(D(0) = d) \quad \text{by the self-motivated treatment assumption} \\
&= P(Y = 0, D = d \mid Z = 0) + P(Y = 1, D = d \mid Z = 0) \quad \text{by consistency} \\
&= q_{d0\cdot 0} + q_{d1\cdot 0}
\end{aligned}
$$

The parameter $\omega_{00}$ can be written

$$
\begin{aligned}
\omega_{00} &= P(D(0) = 0, D(1) = 0) \\
&= P(D(1) = 0)P(D(0) = 0 \mid D(1) = 0) \\
&= P(D(1) = 0) \quad \text{by the self-motivated treatment assumption} \\
&= P(Y = 1, D = 0 \mid Z = 1) + P(Y = 0, D = 0 \mid Z = 1) \quad \text{by consistency} \\
&= q_{01\cdot 1} + q_{00\cdot 1}
\end{aligned}
$$

Finally, since the strata membership probabilities must sum to 1, we have that

$$
\begin{aligned}
\omega_{01} &= P(D(0) = 0, D(1) = 1) \\
&= 1 - (\omega_{00} + \omega_{11} + \omega_{21}) \\
&= 1 - (q_{01\cdot 1} + q_{00\cdot 1} + q_{10\cdot 0} + q_{11\cdot 0} + q_{20\cdot 0} + q_{21\cdot 0})
\end{aligned}
$$

Since the strata membership probabilities $\omega_{d_0 d_1}$ are expressible in terms of the observed quantities $\boldsymbol{q}$, it follows that they are point identified.

We now examine the identifiability of the causal outcome parameters $p_{d_0,d_1\cdot z} = P(Y(z) = 1 \mid D(0) = d_0, D(1) = d_1)$ for $z \in \{0, 1\}$ and $(d_0, d_1) \in \mathcal{D}$, first under Assumptions 1-4 only and then under Assumptions 1-5.

*Identifiability of Causal Outcome Parameters under Assumptions 1-4*

For $d = 1, 2$,

$$
\begin{aligned}
p_{d1\cdot 0} &= \frac{P(Y(0)=1, D(0)=d, D(1)=1)}{P(D(0)=d, D(1)=1)} \\
&= \frac{P(Y(0)=1, D(0)=d)P(D(1)=1 \mid Y(0)=1, D(0)=d)}{\omega_{d1}} \\
&= \frac{q_{d1\cdot 0}}{\omega_{d1}} \text{ (self-motivated treatment and consistency assumptions).}
\end{aligned}
$$

For $d_0 = 0, d_1 = 0$ and $z = 1$, we have

$$
\begin{aligned}
p_{00\cdot 1} &= \frac{P(Y(1)=1, D(0)=0, D(1)=0)}{P(D(0)=0, D(1)=0)} \\
&= \frac{P(Y(1)=1, D(1)=0)P(D(0)=0 \mid Y(1)=1, D(1)=0)}{\omega_{00}} \\
&= \frac{q_{01\cdot 1}}{\omega_{00}} \text{ (self-motivated treatment and consistency assumptions).}
\end{aligned}
$$

Thus, under Assumptions 1-4 the causal outcome parameters $p_{11\cdot 0}, p_{21\cdot 0}$ and $p_{00\cdot 1}$ are point identified. For the remaining causal outcome parameters, we have that

$$
\begin{aligned}
q_{00\cdot 0} &= P(Y=1, D=0 \mid Z=0) \\
&= P(Y(0)=1, D(0)=0) \text{ (consistency assumption)} \\
&= P(Y(0)=1, D(0)=0, D(1)=0) + P(Y(0)=1, D(0)=0, D(1)=1) \\
&= \omega_{00} p_{00\cdot 0} + \omega_{01} p_{01\cdot 0}
\end{aligned}
$$

and

$$
\begin{aligned}
q_{11\cdot 1} &= P(Y=1, D=1 \mid Z=1) \\
&= P(Y(1)=1, D(1)=1) \text{ (consistency assumption)} \\
&= P(Y(1)=1, D(0)=0, D(1)=1) + P(Y(1)=1, D(0)=1, D(1)=1) \\
&\quad + P(Y(1)=1, D(0)=2, D(1)=1) \\
&= \omega_{01} p_{01\cdot 1} + \omega_{11} p_{11\cdot 1} + \omega_{21} p_{21\cdot 1}.
\end{aligned}
$$

Hence, the causal outcome parameters $p_{01\cdot 1}, p_{01\cdot 0}$ and $p_{21\cdot 0}$ are unidentified under

Assumptions 1-4.

*Identifiability of Causal Outcome Parameters under Assumptions 1-5* We now examine the identifiability of the causal outcome parameters under Assumptions 1-5. For $d = 1, 2$, we have that

$$\begin{aligned} p_{d1\cdot 0} &= \frac{P(Y(0) = 1, D(0) = d, D(1) = 1)}{P(D(0) = d, D(1) = 1)} \\ &= \frac{P(Y(0) = 1, D(0) = d)P(D(1) = 1 \mid Y(0) = 1, D(0) = d)}{\omega_{d1}} \\ &= \frac{q_{d1\cdot 0}}{\omega_{d1}} \text{ (self-motivated treatment and consistency assumptions).} \end{aligned}$$

For $d_0, d_1 = 0$ and $z = 1$, we have that

$$\begin{aligned} p_{00\cdot 1} &= \frac{P(Y(1) = 1, D(0) = 0, D(1) = 0}{P(D(0) = 0, D(1) = 0)} \\ &= \frac{P(Y(1) = 1, D(1) = 0)P(D(0) = 0 \mid Y(1) = 1, D(1) = 0)}{P(D(0) = 0, D(1) = 0} \\ &= \frac{q_{01\cdot 1}}{\omega_{00}}. \end{aligned}$$

By Assumption 5 (exclusion restriction), $p_{00\cdot 0} = p_{00\cdot 1}$ and so $p_{00\cdot 0}$ is point identified as well. Finally, since

$$\begin{aligned} P(Y(0) = 1) &= P(Y = 1 \mid Z = 0) \\ &= \omega_{00}p_{00\cdot 0} + \omega_{01}p_{01\cdot 0} + \omega_{11}p_{11\cdot 0} + \omega_{21}p_{21\cdot 0} \end{aligned}$$

we can see by rearranging terms that $p_{11\cdot 0}$ is expressible in terms of point identified quantities and thus is point identified as well. By Assumption 5, $p_{11\cdot 1}$ is point identified as well. Thus, under Assumptions 1-5, the causal outcome parameters $p_{00\cdot 0}, p_{01\cdot 0}, p_{11\cdot 0}, p_{21\cdot 0}, p_{00\cdot 1}$, and $p_{11\cdot 1}$ are point identified.

For $p_{01\cdot 1}$ and $p_{21\cdot 1}$, we can see from the relationship

$$\begin{aligned} q_{11\cdot 1} - q_{11\cdot 0} &= P(Y=1, D=1 \mid Z=1) - P(Y=1, D=1 \mid Z=0) \\ &= P(Y(1)=1, D(1)=1) - P(Y(0)=1, D(0)=1) \\ &= \omega_{01} p_{01\cdot 1} + \omega_{11} p_{11\cdot 1} + \omega_{21} p_{21\cdot 1} - \omega_{11} p_{11\cdot 0} \\ &= \omega_{01} p_{01\cdot 1} + \omega_{21} p_{21\cdot 1} \text{ (exclusion restriction assumption)} \end{aligned}$$

that the causal outcome parameters $p_{01\cdot 1}$ and $p_{21\cdot 1}$ are not point identified.

*Proof of Lemma 3*

Each of the estimators given in this lemma can be written as a function $f$ of the estimators $\hat{\boldsymbol{q}}$, where the function $f$ is given by the identifying equations presented in Lemma 2. By inspecting the right-hand sides of these identifying equations, it is obvious that these are continuous functions. It was established in Lemma 1 that $\hat{\boldsymbol{q}}$ is a $\sqrt{n}$ consistent and asymptotically normal estimator for $\boldsymbol{q}$. Hence, it follows from the Continuous Mapping Theorem and the Delta Method that each of the estimators considered in Lemma 3 must be a $\sqrt{n}$ consistent and asymptotically normal estimator for its respective population parameter.

## *A.2 Functions Related to Large Sample Distributions of Bounds Estimators*

In this appendix, we give detailed expressions for the functions that relate the vector of observed-data quantities $\boldsymbol{q}$ to the expressions for the modified upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ described in Section 1.6. These functions can in turn be used to approximate the asymptotic distribution for the upper and lower bounds. We start with the form of the upper and lower bounds under Assumptions 1-4.

Let $\boldsymbol{g}_{01} : \mathbb{R}^8 \mapsto \mathbb{R}^2$ be the function given by

$$\begin{aligned} \boldsymbol{g}_{01}(x_1, x_2, \ldots, x_7, x_8) &= \begin{pmatrix} g_{01}^1(x_1, x_2, \ldots, x_7, x_8) \\ g_{01}^2(x_1, x_2, \ldots, x_7, x_8) \end{pmatrix} \\ &= \begin{pmatrix} \frac{x_8 - x_1 - x_6 - x_7}{1-(x_2+x_3+x_4+x_5+x_6+x_7)} \\ \frac{x_8 - x_3 - x_2 - x_4 - x_5 - x_7}{1-(x_2+x_3+x_4+x_5+x_6+x_7)} \end{pmatrix} \end{aligned}$$

and let $\boldsymbol{g}_{21} : \mathbb{R}^8 \mapsto \mathbb{R}^2$ be the function given by

$$\begin{aligned}\boldsymbol{g}_{21}(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_7}, \boldsymbol{x_8}) &= \begin{pmatrix} g_{21}^1(x_1, x_2, \ldots, x_7, x_8) \\ g_{21}^2(x_1, x_2, \ldots, x_7, x_8) \end{pmatrix} \\ &= \begin{pmatrix} \frac{x_8 - x_5}{x_4 + x_5} \\ \frac{x_8 - x_3 - x_2 - (1 - \sum_{i=2}^{7} x_i) - x_5}{x_4 + x_5} \end{pmatrix}\end{aligned}$$

If the unrestrained upper and lower bounds on the parameters $p_{01\cdot 1}, p_{01\cdot 0}$ and $p_{21\cdot 1}$ all fall in the interval $[0, 1]$, then evaluating $\boldsymbol{g}_{01}$ at $\boldsymbol{q}$ yields the upper and lower bounds $(\Delta_{01}^U, \Delta_{01}^L)^T$ for $\Delta_{01}$ derived under Assumptions 1-4, while evaluating it at $\hat{\boldsymbol{q}}$ yields the estimators $(\hat{\Delta}_{01}^U, \hat{\Delta}_{01}^L)^T$. Similarly, evaluating $\boldsymbol{g}_{21}$ at $\boldsymbol{q}$ yields the upper and lower bounds $(\Delta_{21}^U, \Delta_{21}^L)^T$ for $\Delta_{21}$ derived under Assumptions 1-4, while evaluating it at $\hat{\boldsymbol{q}}$ yields the estimators $(\hat{\Delta}_{21}^U, \hat{\Delta}_{21}^L)^T$.

We now describe the form of the upper and lower bounds on $\Delta_{01}$ and $\Delta_{21}$ under Assumptions 1-5. Let $\boldsymbol{h}_{01} : \mathbb{R}^8 \mapsto \mathbb{R}^2$ be the function given by

$$\begin{aligned}\boldsymbol{h}_{01}(x_1, x_2, \ldots, x_7, x_8) &= \begin{pmatrix} h_{01}^1(x_1, x_2, \ldots, x_7, x_8) \\ h_{01}^2(x_1, x_2, \ldots, x_7, x_8) \end{pmatrix} \\ &= \begin{pmatrix} \frac{x_8 - x_2 - x_1 + x_7}{1 - (x_2 + x_3 + x_4 + x_5 + x_6 + x_7)} \\ \frac{x_8 - x_2 - x_1 + x_7 - x_4 - x_5}{1 - (x_2 + x_3 + x_4 + x_5 + x_6 + x_7)} \end{pmatrix}\end{aligned}$$

and let $\boldsymbol{h}_{21} : \mathbb{R}^8 \mapsto \mathbb{R}^2$ be the function given by

$$\begin{aligned}\boldsymbol{h}_{21}(x_1, x_2, \ldots, x_7, x_8) &= \begin{pmatrix} h_{21}^1(x_1, x_2, \ldots, x_7, x_8) \\ h_{21}^2(x_1, x_2, \ldots, x_7, x_8) \end{pmatrix} \\ &= \begin{pmatrix} \frac{x_8 - x_2 - x_5}{x_4 + x_5} \\ \frac{x_8 - x_2 - (1 - (x_2 + x_3 + x_4 + x_5 + x_6 + x_7) - x_5}{x_4 + x_5} \end{pmatrix}.\end{aligned}$$

If the unrestrained upper and lower bounds on the parameters $p_{01\cdot 1}$ and $p_{21\cdot 1}$ both fall in the interval $[0, 1]$, then evaluating $\boldsymbol{h}_{01}$ at $\boldsymbol{q}$ yields the upper and lower bounds

$(\Delta_{01}^U, \Delta_{01}^L)^T$ for $\Delta_{01}$ derived under Assumptions 1-5, while evaluating it at $\hat{\boldsymbol{q}}$ yields the estimators $(\hat{\Delta}_{01}^{U,ER}, \hat{\Delta}_{01}^{L,ER})^T$. Similarly, evaluating $\boldsymbol{h}_{21}$ at $\boldsymbol{q}$ yields the upper and lower bounds $(\Delta_{21}^U, \Delta_{21}^L)^T$) for $\Delta_{21}$ derived under Assumptions 1-5, while evaluating it at $\hat{\boldsymbol{q}}$ yields the estimators $(\hat{\Delta}_{21}^{U,ER}, \hat{\Delta}_{21}^{L,ER})^T$.

# Appendix B

*Proof of Lemma 6.*

*Consistency.*

The estimators $\hat{q}_{dy\cdot zk}$ can be written as a ratio of sample means

$$\hat{q}_{dy\cdot zk} = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i = y)\mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)\mathbb{1}(S_i = s)}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(S_i = s)}.$$

By the Weak Law of Large Numbers (WLLN), the numerator converges in probability to $E[\mathbb{1}(Y_i = y)\mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)\mathbb{1}(S_i = s)] = P(Y = y, D = d, Z = z, S = s)$. Similarly, by the WLLN the denominator converges in probability to $E[\mathbb{1}(Z_i = z)\mathbb{1}(S_i = s)] = P(Z = z, S = s)$. Thus, by Slutsky's Theorem and the Continuous Mapping Theorem we have that

$$\begin{aligned}\hat{q}_{dy\cdot zk} &\xrightarrow{p} \frac{P(Y = y, D = d, Z = z, S = s)}{P(Z = z, S = s)} \\ &= P(Y = y, D = d \mid Z = z, S = s). \qquad \blacksquare\end{aligned}$$

*Asymptotic Normality.*

By the assumption of independence between sites, we can consider the asymptotic behavior of the observed-data vectors $\boldsymbol{q}_k$ separately. We let $\hat{p}_{1k} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = 1)\mathbb{1}(S_i = k)$ and $\hat{p}_{0k} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = 0)\mathbb{1}(S_i = k)$. For the $k^{th}$ site, the quantities

$\hat{q}_{dy\cdot zk} - q_{dy\cdot zk}$ can be written

$$\begin{aligned}
\hat{q}_{dy\cdot z} - q_{dy\cdot z} &= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D_i = d)\mathbb{1}(Z_i = z)\mathbb{1}(Y_i = y)\mathbb{1}(S_i = k)}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(S_i = k)} - q_{dy\cdot zk} \\
&= (\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(S_i = s))^{-1}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(D_i = d)\mathbb{1}(Y_i = y)\mathbb{1}(Z_i = z)\mathbb{1}(S_i = k) - q_{dy\cdot zk} \\
&= (\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = z)\mathbb{1}(S_i = k))^{-1}\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}(D_i = d)\mathbb{1}(Y_i = y) - q_{dy\cdot zk})\mathbb{1}(Z_i = z)\mathbb{1}(S_i = k) \\
&\equiv (\hat{p}_{zk})^{-1}\tilde{q}_{dy\cdot zk},
\end{aligned}$$

and thus the vector $\hat{\boldsymbol{q}}_k - \boldsymbol{q}_k$ can be written

$$\hat{\boldsymbol{q}}_k - \boldsymbol{q}_k = \hat{\boldsymbol{Z}}_k^{-1}\tilde{\boldsymbol{q}}_k$$

where $\hat{\boldsymbol{Z}}_k$ is a $8 \times 8$ diagonal matrix with entries

$$\hat{\boldsymbol{Z}} = \begin{pmatrix}
\hat{p}_{0k} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \hat{p}_{0k} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \hat{p}_{0k} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \hat{p}_{0k} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \hat{p}_{0k} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \hat{p}_{1k} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \hat{p}_{1k} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{p}_{1k}
\end{pmatrix}$$

.

The elements of $\tilde{\boldsymbol{q}}_k$ are sample means of the centered random variables $X_{dy\cdot zk} = (\mathbb{1}(D = d)\mathbb{1}(Y = y) - q_{dy\cdot zk})\mathbb{1}(S_i = k)\mathbb{1}(Z = z)$ which have expectation and variance

$$E[X_{dy\cdot zk}] = 0$$

$$Var[X_{dy\cdot zk}] = p_{zk}q_{dy\cdot zk}(1 - q_{dy\cdot zk}).$$

For $d, y, d', y'$ where either $d \neq d'$, $y \neq y'$ or both, the covariance is given by

$$Cov(X_{dy \cdot zk}, X_{d'y' \cdot zk}) = -p_z q_{dy \cdot zk} q_{d'y' \cdot zk}.$$

By the CLT,

$$\sqrt{n}\tilde{\boldsymbol{q}}_k \xrightarrow{d} N_8(\mathbf{0}, \Sigma^*_{qk})$$

where

$$\Sigma^*_q = \begin{pmatrix} p_{0k}\Sigma_{0k} & \mathbf{0} \\ \mathbf{0} & p_{1k}\Sigma_{1k} \end{pmatrix}.$$

By the LLN

$$\hat{p}_{zk} \xrightarrow{p} p_{zk}$$

and thus by the Continuous Mapping Theorem

$$\hat{\boldsymbol{Z}}_k^{-1} \xrightarrow{p} \boldsymbol{Z}_k^{-1}$$

where

$$\boldsymbol{Z}_k = \begin{pmatrix} p_{0k} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{0k} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{0k} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{0k} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{0k} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{1k} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{1k} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{1k} \end{pmatrix}.$$

We can therefore conclude that

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{q}}_k - \boldsymbol{q}_k) &= \hat{\boldsymbol{Z}}_k^{-1}\sqrt{n}\tilde{\boldsymbol{q}}_k \\ &\xrightarrow{d} N_8(\boldsymbol{0}, \boldsymbol{Z}_k^{-1}\Sigma_q^*\boldsymbol{Z}_k^{-1}) \\ &= N_8(\boldsymbol{0}, \Sigma_{qk})\end{aligned}$$

by Slutsky's theorem. Thus, by the independence of the data from the $K$ sites it follows that

$$\sqrt{n}(\hat{\boldsymbol{q}} - \boldsymbol{q}) \xrightarrow{d} N_{8K}(\boldsymbol{0}, \Sigma_{\boldsymbol{q}}). \qquad \blacksquare$$

# Appendix C

| Binary outcome<br>Assumptions | Parameters | Hyperparameters |
|---|---|---|
| ER, SA | $(\omega_c, \omega_n, p_{c0}, p_{c1}, p_n)$ | $(\gamma_c, \gamma_n, \alpha_{c0}, \beta_{c0}, \alpha_{c1}, \beta_{c1}, \alpha_n, \beta_n)$ |
| ER, No SA | $(\omega_c, \omega_n, \omega_a, p_{c0}, p_{c1}, p_n, p_a)$ | $(\gamma_c, \gamma_n, \gamma_a, \alpha_{c0}, \beta_{c0}, \alpha_{c1}, \beta_{c1}, \alpha_n, \beta_n, \alpha_a, \beta_a)$ |
| No ER, SA | $(\omega_c, \omega_n, p_{c0}, p_{c1}, p_{n0}, p_{n1})$ | $(\gamma_c, \gamma_n, \alpha_{c0}, \beta_{c0}, \alpha_{c1}, \beta_{c1}, \alpha_{n0}, \beta_{n0}, \alpha_{n1}, \beta_{n1})$ |
| No ER, No SA | $(\omega_c, \omega_n, \omega_a, p_{c0}, p_{c1}, p_{n0}, p_{n1}, p_{a0}, p_{a1})$ | $(\gamma_c, \gamma_n, \gamma_a, \alpha_{c0}, \beta_{c0}, \alpha_{c1}, \beta_{c1}, \alpha_{n0}, \beta_{n0}, \alpha_{n1}, \beta_{n1}, \alpha_{a0}, \beta_{a0}, \alpha_{a1}, \beta_{a1})$ |

| Normal outcome<br>Assumptions | Parameters | Hyperparameters |
|---|---|---|
| ER, SA | $(\omega_c, \omega_n, \mu_{c0}, \sigma^2_{c0}, \mu_{c1}, \sigma^2_{c1}, \mu_n, \sigma^2_n)$ | $(\gamma_c, \gamma_n, \theta_{c0}, \tau_{c0}, a_{c0}, b_{c0}, \theta_{c1}, \tau_{c1}, a_{c1}, b_{c1}, \theta_n, \tau_n, a_n, b_n)$ |
| ER, No SA | $(\omega_c, \omega_n, \omega_a, \mu_{c0}, \sigma^2_{c0}, \mu_{c1}, \sigma^2_{c1}, \mu_n, \sigma^2_n, \mu_a, \sigma^2_a)$ | $(\gamma_c, \gamma_n, \gamma_a, \theta_{c0}, \tau_{c0}, a_{c0}, b_{c0}, \theta_{c1}, \tau_{c1}, a_{c1}, b_{c1}, \theta_n, \tau_n, a_n, b_n, \theta_a, \tau_a, a_a, b_a)$ |
| No ER, SA | $(\omega_c, \omega_n, \mu_{c0}, \sigma^2_{c0}, \mu_{c1}, \sigma^2_{c1}, \mu_{n0}, \sigma^2_{n0}, \mu_{n1}, \sigma^2_{n1})$ | $(\gamma_c, \gamma_n, \theta_{c0}, \tau_{c0}, a_{c0}, b_{c0}, \theta_{c1}, \tau_{c1}, a_{c1}, b_{c1}, \theta_{n0}, \tau_{n0}, a_{n0}, b_{n0}, \theta_{n1}, \tau_{n1}, a_{n1}, b_{n1})$ |
| No ER, No SA | $(\omega_c, \omega_n, \omega_a, \mu_{c0}, \sigma^2_{c0}, \mu_{c1}, \sigma^2_{c1}, \mu_{n0}, \sigma^2_{n0}, \mu_{n1}, \sigma^2_{n1}, \mu_{a0}, \sigma^2_{a0}, \mu_{a1}, \sigma^2_{a1})$ | $(\gamma_c, \gamma_n, \gamma_a, \theta_{c0}, \tau_{c0}, a_{c0}, b_{c0}, \theta_{c1}, \tau_{c1}, a_{c1}, b_{c1}, \theta_{n0}, \tau_{n0}, a_{n0}, b_{n0}, \theta_{n1}, \tau_{n1}, a_{n1}, b_{n1}, \theta_{a0}, \tau_{a0}, a_{a0}, b_{a0}, \theta_{a1}, \tau_{a1}, a_{a1}, b_{a1})$ |

Table C.1: Complete lists of parameters and hyperparameters for the binary and Normal outcome models under four sets of assumptions. ER = Exclusion Restriction assumption, SA = Strong Access Monotonicity assumption

# Appendix D

Table D.1: Characteristics of the review studies.

| Author (Year) | Program | Delivery Environment | Treatment type | Mean Age in Months (Range) | Group/ One-on-one | Parent Component |
|---|---|---|---|---|---|---|
| Eapen (2013) | Group-based adaptation of ESDM | Community care center | NDBI | 49.6 (36-58) | Primarily group-based | No |
| Ingersoll (2013) | Parent-training program based on Project ImPact | Home | NDBI | 44.9 (26-70) | One-on-one | Yes |
| Kasari (2014) | JASPER | Home | Joint attention skills | 42.3 (NEI) | One-on-one | Yes |
| Lawton (2012) | Teacher-implemented JASPER | School | Joint attention skills | 44.7 (NEI) | One-on-one | No |
| Peters-Scheffer (2010) | Intervention based on Lovaas (2003) | School | One-on-one discrete trial training | 53.1 (38-75) | One-on-one | Yes |
| Salt (2002) | Scottish Centre for Autism treatment | Community care center | Social-developmental | 50.0 (NEI) | Primarily one-on-one | Yes |
| Smith (2010) | Nova Scotia Early Intensive Behavior Intervention Model | School and home | Pivotal Response Treatment | 49.6 (25-72) | One-on-one | Yes |
| Stadnick (2015) | Project imPACT | Community care center | NDBI | 54.9 (18-108) | One-on-one | Yes |
| Vivanti (2014) | Group-based ESDM | Community care center | NDBI | 41.1 (NEI) | Group | Yes |
| Wetherby (2014) | SCERTS | Home | NDBI | 19.6 (NEI) | One-on-one | Yes |