Methods for Estimating Causal Effects of Treatment in Randomized Controlled Trials with Simultaneous Provider and Subject Noncompliance

Elisa Sheng

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Xiao-Hua (Andrew) Zhou, Chair

Thomas Richardson

Susan Shortreed

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Methods for Estimating Causal Effects of Treatment in Randomized Controlled Trials with Simultaneous Provider and Subject Noncompliance

Elisa Sheng

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

Subject noncompliance is a common problem in the analysis of randomized controlled trials (RCTs); with cognitive behavioral interventions, the addition of provider noncompliance further complicates making causal inference. As a motivating example, we consider a RCT of a Motivational Interviewing (MI)-based behavioral intervention for treating problem drug use. Treatment receipt depends on compliance of both a therapist (provider) and a patient (subject) where MI is 'received' when the therapist adheres to the MI protocol and the patient actively participates in the intervention. However, therapists cannot be forced to follow protocol and patients cannot be forced to cooperate in an intervention. In this dissertation, we define causal estimands of interest based on a principal stratification framework, propose methods for estimating these causal estimands, and apply our proposals to a RCT of MI.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank everyone who has made this dissertation possible. I thank my adviser, Andrew Zhou, for his guidance and mentorship throughout my entire graduate student career. I thank David Atkins for helping to motivate the research, as well as his support and encouragement. I would also like to acknowledge and thank my committee members, Thomas Richardson, Susan Shortreed, Dave and Ann Vander Stoep. Finally, I thank the UW Department of Biostatistics and the National Institute of Mental Health for supporting this research through the Biostatistics Training Grant (T32 MH 73521-7).

DEDICATION

to my family

Chapter 0

INTRODUCTION

Statistical analyses of Randomized Controlled Trials (RCTs) typically adhere to the intent-to-treat (ITT) principle, in which data are analyzed according to the subject's treatment assignment, regardless of whether they actually received treatment. However, with subject noncompliance, the ITT effect, which represents the causal effect of treatment *assignment* on subject outcomes, is not the same as the causal effect of treatment (the causal effect of treatment *receipt* on subject outcomes). Furthermore, treatment receipt may depend on compliance of both a provider (e.g. doctors, teachers) and a subject (e.g. patients, students).

As a specific example, consider cognitive behavioral interventions. Compliance of the therapist in addition to the patient is typically problematic. Unlike medication, which is regulated by law and homogenized by manufacturing processes, a behavioral intervention inherently differs from provider to provider. Even when the intervention has a rigorous protocol, therapists cannot be forced to follow it, and hence therapist noncompliance occurs. Furthermore, patient noncompliance occurs because patients cannot be forced to participate in treatment. The focus of this research is to address noncompliance that may occur at two levels: (1) noncompliance of the provider, and (2) noncompliance of the subject level.

Throughout the dissertation we consider a study of Motivational Interviewing (MI), a behavioral intervention that seeks to encourage a patient's own motivation to change a problematic behavior [Krupski et al., 2012]. According to the conceptual model of MI [Miller and Rose, 2010], to affect behavior, the therapist must meet MI proficiency standards (i.e. follow MI protocol) and the patient must become motivated to change (i.e. 'take' the MI). To assess the effect of MI on patient outcomes, researchers would like to compare patient

outcomes under the following scenarios:

(1) the therapist follows MI protocol and patient expresses intent to change; and

(2) the therapist does not follow MI protocol and the patient does not express intent to change.

Clearly assignment to MI does not guarantee either of these conditions hold. Hence the usual ITT analyses are not appropriate for assessing the causal effect of MI, yet statistical methods for making causal inference in such a setting are highly limited.

We develop methodology for estimating causal effects of treatment in the presence of simultaneous provider and subject noncompliance in three phases. In Chapter 1 we develop methodology that assumes provider-subject pairs are independent and identically distributed (iid). In Chapter 2 we extend the methodology to estimate conditional causal effects, relaxing the iid assumption to hold conditionally. In Chapter 3 we extend the methodology to estimate conditional causal effects in clustered RCTs, relaxing the iid assumption further and more closely resembling the study design of Krupski et al. [2012] where patients are nested within therapists. At each phase we illustrate the methods on data from the RCT of MI described in Krupski et al. [2012]. In Chapter 4 we consider ways in which the proposed methods may be useful, and areas of research left open.

Chapter 1

ESTIMATING CAUSAL EFFECTS OF TREATMENT IN RANDOMIZED CONTROLLED TRIALS

In the causal inference literature, several methods have been proposed to address subject-only (one-level) noncompliance, however few have considered simultaneous subject and provider (two-level) noncompliance. Potential outcomes [Neyman, 1990, Rubin, 1974, 1978] and principal stratification [Frangakis and Rubin, 2002] are typically used to define causal estimands of interest. The three general approaches to estimating these causal estimands are: the method of moments approach, (e.g. Angrist et al. [1996], Frangakis and Rubin [1999], Yau and Little [2001], Albert [2002], Levy et al. [2004], Zhou and Li [2006], Taylor and Zhou [2009a], Ding et al. [2012]); the maximum likelihood approach (e.g. Cuzick et al. [1997], Peng et al. [2004], O'Malley and Normand [2005], Zhou and Li [2006], Shepherd et al. [2006], Jo et al. [2008]), and the Bayesian approach (e.g. Hirano et al. [2000], Frangakis et al. [2002], Peng et al. [2004], Barnard et al. [2003], Richardson et al. [2011]). The vast majority of proposals consider the one-level noncompliance setting (e.g. assume perfect provider compliance or define "compliance" to mean compliance of both provider and subject). There is one exception: Schochet and Chiang [2011] uses a method of moments (MOM) approach to estimate a proposed causal estimand for the two-level noncompliance setting.

Similar to the proposal of Schochet and Chiang [2011], we use a principal stratification framework to define a causal estimand of interest when both provider and subject compliance underlie the definition of treatment receipt. We call the estimand of interest the *average causal effect of treatment among provider-subject pairs that comply with assignment* and denote it by $ACE(cc)$. In this chapter we examine all possible sets of assumptions for identifying $ACE(cc)$, whereas Schochet and Chiang [2011] only considered one possibility.

We also provide tests to validate assumptions. We propose MOM and maximum likelihood (ML) estimators of $ACE(cc)$ under newly proposed identifying assumptions, extending the existing statistical methodology for making causal inference in the presence of both provider and subject noncompliance.

1.1 Definition of the Causal Effect of Treatment

To motivate the causal estimand of interest, first consider the ideal setting: perfect provider and subject compliance. In this setting, the ITT effect is equivalent to the causal effect of treatment. If the provider-subject pair is assigned to treatment then the provider follows the treatment protocol and the subject receives the treatment, otherwise treatment protocol is not followed and treatment is not received. Standard ITT analyses (e.g. a 2-sample t-test) can be used to estimate the causal effect of treatment. Next consider the one-level noncompliance setting; we still have perfect provider compliance, but subjects in either arm of the study may receive the opposite of their assignment. The causal effect of treatment estimated by an ITT analysis is then biased toward zero. A solution is to instead consider the Complier Average Causal Effect of treatment, which is the ITT effect among the subgroup of subjects that take treatment when it is assigned and do not take treatment when it is not assigned, or *compliers*.[1] By restricting attention to compliers, the ITT effect is the causal effect of treatment. Now turning to the setting with two-level noncompliance, whether subjects comply with treatment assignment is complicated by the fact that now providers may not comply with the assignment. Nevertheless, by taking a principal stratification approach we can extend the notion of the Complier Average Causal Effect to the two-level noncompliance setting.

First, we need to define some notation. Suppose there are n independent provider-subject pairs and let i index the pairs. Denote treatment assignment for the ith pair by Z_i ($Z_i = 1$ if the pair is assigned to treatment and $Z_i = 0$ otherwise). For the ith pair, denote the

[1]Note that whether the subject is a complier cannot be identified from the observed data without further assumptions.

provider's observed compliance by D_i^{P} and the subject's observed compliance by D_i^{S}; $D_i^{\mathrm{P}} = 1$ if the provider follows the treatment protocol and $D_i^{\mathrm{P}} = 0$ otherwise; $D_i^{\mathrm{S}} = 1$ if the subject takes the treatment and $D_i^{\mathrm{S}} = 0$ otherwise. Denote a binary observed outcome for the subject in the ith pair by Y_i. Let $\mathcal{N}$ denote the set of all n-dimensional column vectors of zeros and ones indicating all possible treatment assigments; hence $|\mathcal{N}| = 2^n$ and $\mathbf{z} \in \mathcal{N}$ represents one possible treatment assignment to the n provider-subject pairs. For the ith provider-subject pair, define the potential compliance under $\mathbf{z}$ by $D_i^{\mathrm{P}}(\mathbf{z})$ and $D_i^{\mathrm{S}}(\mathbf{z})$ for $\mathbf{z} \in \mathcal{N}$, and define the potential subject outcome under assignment $\mathbf{z}$ by $Y_i(\mathbf{z}) \equiv Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z}))$ for $\mathbf{z} \in \mathcal{N}$ where $\mathbf{D}^{\mathrm{P}}(\mathbf{z}) = (D_1^{\mathrm{P}}(\mathbf{z}), ..., D_n^{\mathrm{P}}(\mathbf{z}))$ and $\mathbf{D}^{\mathrm{S}}(\mathbf{z}) = (D_1^{\mathrm{S}}(\mathbf{z}), ..., D_n^{\mathrm{S}}(\mathbf{z}))$.

There are 16 subgroups of provider-subject pairs defined by contrasting potential compliance under $z_i = 1$ (treatment) and $z_i = 0$ (no treatment) with $\mathbf{z}_{-i}$ otherwise fixed, illustrated in Figure 1.1. Each subgroup represents a *compliance type* that can be summarized by a two letter combination. The first letter corresponds to the provider's potential compliance, and the second to the subject's. For example, for type na, the provider *never* follows treatment protocol, and the patient *always* takes the treatment, regardless of assignment. Similarly, following the usual naming convention in the causal inference literature, the letters a, n, c and d correspond to $always$, $never$, $complies$ and $defies$. A provider/subject 'complies' when they follow treatment protocol/take treatment only if it is assigned, and 'defies' if they do the opposite. Denote the compliance type of the ith provider-subject pair by C_i, which takes on values in $\mathcal{C} = \{st : s, t \in \{c, a, n, d\}\}$.

Recall that comparing subject outcomes under treatment receipt and non-receipt is of particular interest to researchers. This comparison can be described by the following scenarios: (1) treatment is assigned, the provider follows the treatment protocol and the subject takes treatment and (2) treatment is not assigned, the provider does not follow the treatment protocol and the subject does not take the treatment. The difference in expected potential outcomes for provider-subject pairs with compliance type cc defines the causal

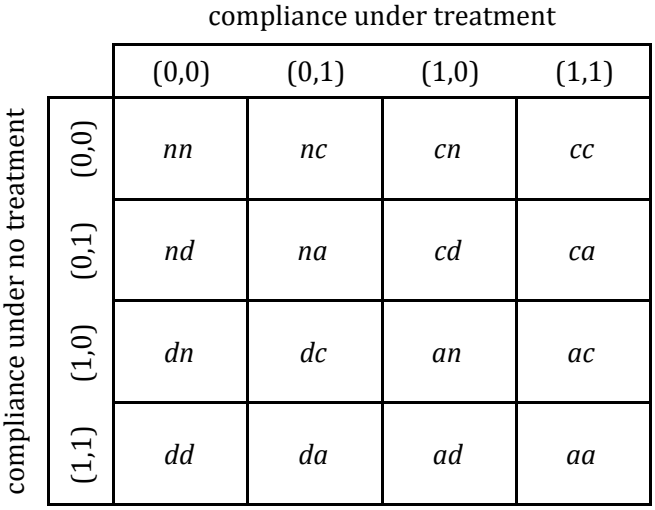


Figure 1.1: Illustration of compliance types where potential compliance has both a provider and subject component. The pair of values $(D_i^P(\mathbf{z}), D_i^S(\mathbf{z})) \in \{0,1\} \times \{0,1\}$ denote the provider and subject's potential compliance, respectively, under treatment ($\mathbf{z} = (z_1, ..., 1, ..., z_n)$) and no treatment ($\mathbf{z} = (z_1, ..., 0, ..., z_n)$). Zero indicates the subject/provider complies with treatment, one indicates noncompliance.

effect of treatment under assignment $\mathbf{z}$:

$$P(Y_i(Z_1 = z_1, ..., Z_i = 1, ..., Z_n = z_n) = 1 \mid C_i = cc) \\ - P(Y_i(Z_1 = z_1, ..., Z_i = 0, ..., Z_n = z_n) = 1 \mid C_i = cc). \quad (1.1)$$

However, we are not able to observe both potential outcomes (nor the compliance type); this is known as the fundamental problem of causal inference [Holland, 1986]. Furthermore, (1.1) depends on treatment assignment of every other provider-subject pair, allowing for arbitrary interference, which is not particularly relevant to individual-level behavioral interventions. In the following section, we impose assumptions in order to define a causal estimand of interest and identify this estimand.

1.2 A Causal Estimand of Interest: $ACE(cc)$

We first require three assumptions to define a causal estimand of interest: Randomization, the Stable Unit Treatment Value Assumption (SUTVA), and Independent and Identical Distribution (iid) of the provider-subject pairs.

Assumption 1.1 (Randomization) *For $i = 1, ..., n$,*

$$Z_i \perp\!\!\!\perp D_i^{\mathrm{P}}(\mathbf{z}), D_i^{\mathrm{S}}(\mathbf{z}), Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})) \text{ for } \mathbf{z} \in \mathcal{N}.$$

In words, we assume that treatment assignment does not depend on provider or subject potential compliance behavior nor potential outcomes. This assumption could be violated, for example, if patients that were deemed unlikely to cooperate in therapy were assigned to no treatment.

Assumption 1.2 (SUTVA) *For all $\mathbf{z}$ and $i = 1, ..., n$,*

$$D_i^{\mathrm{P}}(\mathbf{z}) = D_i^{\mathrm{P}}(z_i), D_i^{\mathrm{S}}(\mathbf{z}) = D_i^{\mathrm{S}}(z_i) \text{ and } Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})) = Y_i(z_i, D_i^{\mathrm{P}}(z_i), D_i^{\mathrm{S}}(z_i)).$$

Also, the observed data $(Y_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}})$ are related to the potential values as follows:

$$D_i^{\mathrm{P}} = \sum_{z \in \{0,1\}} D_i^{\mathrm{P}}(z) \mathbb{1}(Z_i = z), \; D_i^{\mathrm{S}} = \sum_{z \in \{0,1\}} D_i^{\mathrm{S}}(z) \mathbb{1}(Z_i = z), \quad \textit{and}$$

$$Y_i = \sum_{z\in\{0,1\}} \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} Y_i(z,u,v)\mathbb{1}(Z_i = z, D_i^{\mathrm{P}}(z) = u, D_i^{\mathrm{S}}(z) = v).$$

SUTVA is comprised of two assumptions: (1) each provider-subject pair is not influenced by other provider-subject pairs, or *no interference* and (2) the observed data are consistent with potential values, or *consistency*. Interference could occur in a setting where patients are in contact with each other, e.g. a group therapy setting. Consistency could be violated if the treatment has multiple versions. For example, with a continuous outcome we might expect small differences among MI treatments that range from minimally to highly proficient.[2]

Assumption 1.3 (iid) *For* $i = 1, ..., n$,

$$Y_i \mid Z_i = z, C_i = st \overset{iid}{\sim} Bernoulli(\eta_{stz}) \text{ for } st \in \mathcal{C} \text{ and } z \in \{0,1\} \text{ and}$$

$$(C_1, ..., C_n) \sim Multinomial(n, \pi = (\pi_{st} : st \in \mathcal{C})).$$

Each provider-subject pair is assumed to represent one observation from the same distribution. This assumption could be violated if provider-subject pairs behave differently conditional on covariates (which will be addressed in Chapter 2) or subjects are nested within providers and hence have correlated compliance/outcomes within providers (which will be addressed in Chapter 3).

Under Assumptions 1.1 - 1.3, define the causal estimand of interest, *the average causal effect of treatment among provider-subject pairs that comply with assignment* denoted by $ACE(cc)$, as follows.

Definition 1.1 ($ACE(cc)$)

$$\begin{aligned} ACE(cc) &= P(Y_i = 1 \mid Z_i = 1, C_i = cc) - P(Y_i = 1 \mid Z_i = 0, C_i = cc) \\ &= \eta_{cc1} - \eta_{cc0}. \end{aligned}$$

[2]Since the outcome is binary, it is unlikely that differences in MI above and beyond proficiency would affect the outcome.

The statistical model under Assumptions 1.1 - 1.3 is illustrated in Figure 1.2, inspired by the figures of Richardson et al. [2011]. Since the variables $Z_i, C_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i$ are all discrete, the unobserved and observed data distributions can be defined by multinomial distributions. The unobserved data distribution is multinomial distribution with 32 parameters; the circular nodes of Figure 1.2 depict the parameters of this distribution. Similarly, the rectangular nodes depict the parameters of the observed data multinomial distribution. The arrows indicate which parameters of the unobserved data distribution impact the parameters of the observed data distribution. For example,

$$P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) = \eta_{cn0}\pi_{cn} + \eta_{nn0}\pi_{nn} + \eta_{nc0}\pi_{nc} + \eta_{cc0}\pi_{cc}$$

since provider-subject pairs observed to have $Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0$ are of type cn, nn, nc or cc (by the law of total probability). The parameters of interest (orange nodes) cannot be identified from the observed data because there is no way to distinguish η_{cn0}, η_{nn0}, η_{nc0} from η_{cc0} in $P(Y_i \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0)$ nor η_{aa1}, η_{ca1}, η_{ac1} from η_{cc1} in $P(Y_i \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1)$. We will now consider additional assumptions that identify $ACE(cc)$.

Analogous to assumptions made to identify the Complier Average Causal Effect in the one-level noncompliance setting, we now consider a Monotonicity assumption.

Assumption 1.4 (Monotonicity) *For $i = 1, ..., n$,*

$$D_i^{\mathrm{P}}(1) \geq D_i^{\mathrm{P}}(0) \quad \textit{and} \quad D_i^{\mathrm{S}}(1) \geq D_i^{\mathrm{S}}(0).$$

That is, no provider-subject pairs have compliance type $dc, da, dn, dd, cd, ad,$ or nd.

Assumption 1.4 presumes no provider or subject defies treatment assignment. In behavioral interventions, this would rule out three kinds of therapist-patient pairs. First, pairs where the therapist would follow the treatment protocol if $z = 0$, but would not follow protocol if $z = 1$ (types dc, da, dn and dd). This scenario is implausible as it would suggest therapists intentionally do the opposite of what they are assigned in order to undermine the trial. Second, pairs where the patient would not participate in treatment if $z = 1$ *and* the therapist

followed the treatment protocol, but would participate if $z = 0$ *and* the therapist did not follow protocol (type *cd*). Third, pairs where patients participate under $z = 0$ but not under $z = 1$ and the therapist always/never follows the treatment protocol (types *ad* and *nd*). These last two scenarios are implausible since treatment assignment is unlikely to be divulged to the subject as blinding patients to treatment assignment is the gold standard. However, an undermining therapist or an informed consent clause that effectively unblinds patients are examples of when this assumption could be violated. The model illustrated in Figure 1.2 is simplified under the addition of Assumption 1.4 as parameters corresponding to compliance types dc, da, dn, dd, cd, ad and nd are eliminated.

Intuitively, there should be no direct effect of treatment assignment on subject compliance other that through provider compliance, which motivates our next assumption.

Assumption 1.5 (Additional Compliance Type Restrictions) *For* $i = 1, ..., n$,

$$\textit{if} \;\; D_i^{\mathrm{P}}(1) = D_i^{\mathrm{P}}(0) \;\; \textit{then} \;\; D_i^{\mathrm{S}}(1) = D_i^{\mathrm{S}}(0).$$

That is, no provider-subject pairs have compliance type ac or nc.

In words, Assumption 1.5 presumes that subjects do not vary in compliance across treatment assignment when provider compliance to treatment protocol is not affected by treatment assignment. This scenario is unlikely to be violated since treatment assignment is unlikely to be divulged to the subject but could be violated if the subject were to become unblinded. Together, Assumptions 1.4 and 1.5 rule out 9 of the 16 compliance types, leaving 7 principal strata. Under the assumptions listed so far, the distribution over compliance types, $\pi = (\pi_{cc}, \pi_{cn}, \pi_{ca}, \pi_{nn}, \pi_{na}, \pi_{an}, \pi_{aa})$, is identified (shown in Section A.1 of Appendix A).

Now consider a straightforward extension of the Exclusion Restriction assumption in the one-level noncompliance setting defined by Angrist et al. [1996].

Assumption 1.6 (Stochastic Exclusion Restrictions)

$$\eta_{st1} = \eta_{st0} \equiv \eta_{st} \textit{ for } st \in \{aa, an, na, nn\}.$$

Assumption 1.6 presumes the distribution of patient outcomes does not change with treatment assignment when the provider follows (or does not follow) treatment protocol regardless of assignment. Again, this scenario is unlikely to be violated since treatment assignment is unlikely to be divulged to the subject but could be violated if the subject were to become unblinded.

Despite restricting to 7 of the 16 compliance types with Assumptions 1.4 and 1.5 and placing restrictions on conditional expected outcomes in Assumption 1.6, Assumptions 1.1 - 1.6 are still not enough to identify $ACE(cc)$ from the observed data, as shown in Figure 1.3. From this figure, we have

$$P(Y_i = 0 \mid Z_i = 0, D^{\mathrm{P}} = 0, D^{\mathrm{S}} = 0) = \eta_{cn0}\pi_{cn} + \eta_{nn}\pi_{nn} + \eta_{cc0}\pi_{cc}$$

where $\eta_{nn}, \pi_{cn}, \pi_{nn}$ and π_{cc} are identified from the observed data:

$$\begin{aligned}
\eta_{nn} &= P(Y_i = 1 \mid Z_i = 1, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
\pi_{cn} &= P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 1) - P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) \\
\pi_{nn} &= P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1) \\
\pi_{cc} &= P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) - \pi_{cn} - \pi_{nn}
\end{aligned}$$

but η_{cc0} (and η_{cn0}) cannot be identified. Similarly, η_{cc1} (and η_{ca1}) cannot be identified. However, by inspection, identification of $ACE(cc)$ can be achieved by making additional stochastic exclusion restriction-type assumptions. Specifically, let $(\eta_{cn0}, \eta_{ca1}) = (r_0, r_1) \in \mathcal{R}$ where

$$\begin{aligned}
\mathcal{R} = \{(r_0, r_1) \;\; : \;\; & r_0 \in \{\eta_{nn}, \eta_{an}, \eta_{cn1}, \eta_{aa}, \eta_{na}, \eta_{ca0}, \eta_{cn1}\}, \\
& r_1 \in \{\eta_{nn}, \eta_{an}, \eta_{cn1}, \eta_{aa}, \eta_{na}, \eta_{ca0}, \eta_{cn0}\}\} \;/\; \{(\eta_{ca1}, \eta_{cn0})\}.
\end{aligned}$$

Not all of the $|\mathcal{R}| = 48$ possibilities are scientifically sensible. For example, it is difficult to believe that there are examples where $\eta_{cn0} = \eta_{aa}$. In other words, the expected outcome for a patient who never takes treatment is the same as that of a patient who always accepts

treatment. We now give some scientifically plausible additional exclusion restrictions which allow $ACE(cc)$ to be identified.

One option for achieving identification of $ACE(cc)$ is to assume if the ith provider-subject pair is of compliance type $st \neq cc$, then

$$P(Y_i = 1 \mid Z_i = 1, C_i = st) = P(Y_i = 1 \mid Z_i = 0, C_i = st).$$

More formally, the following assumption, in addition to Assumptions 1.1 - 1.6 holds.

Assumption 1.7 (Additional Stochastic Exclusion Restrictions (option A))

$$\eta_{cn0} = \eta_{cn1} \quad and \quad \eta_{ca1} = \eta_{ca0}.$$

Proposition 1.1 *Under Assumptions 1.1 - 1.6 and 1.7, $ACE(cc)$ is identifiable.*

Proposition 1.1 is identical to the proposal in Schochet and Chiang [2011] and the identification result is illustrated in Figure 1.4. That is,

$$\eta_{cc0} = P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) - \eta_{cn} - \eta_{nn}$$

where

$$\begin{aligned}
\eta_{cn} &= \frac{1}{\pi_{cn}} P(Y_i = 1 \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1) - \pi_{an}\eta_{an}, \\
\eta_{an} &= P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0), \\
\pi_{an} &= P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 0),
\end{aligned}$$

and π_{cn} as defined previously. In other words, Assumption 1.7 allows the previously unidentified parameter η_{cn0} to be identified from the equation $\eta_{cn0} = \eta_{cn1} \equiv \eta_{cn}$, which leads to the identification of η_{cc0}. However in behavioral interventions, Assumption 1.7 seems implausible. For example, when a patient never participates in treatment whether $z = 0$ or $z = 1$, but the therapist performs differently depending on z, it is likely that the intervention under $z = 1$ would have a different effect on the patient's outcome compared to the intervention under $z = 0$.

An alternative option for achieving identification of $ACE(cc)$ is to assume 1.1 - 1.6 and that the distribution of patient outcomes are the same among therapist-patient pairs with the same potential compliance behavior.

Assumption 1.8 (Additional Stochastic Exclusion Restrictions (option B))

$$\eta_{cn0} = \eta_{nn} \equiv \eta_{\bullet n} \quad and \quad \eta_{ca1} = \eta_{aa} \equiv \eta_{\bullet a}.$$

The "$\bullet$" in $\eta_{\bullet n}$ and $\eta_{\bullet a}$ denotes the provider-specific contribution to the compliance type varies (e.g. first letter of the compliance type is c or n) but the subject-specific contribution does not (e.g. second letter is n).

Proposition 1.2 *Under Assumptions 1.1 - 1.6 and 1.8, $ACE(cc)$ is identifiable.*

The proof of Proposition 1.2 is given in Appendix A. As illustrated in Figure 1.5, we have

$$\begin{aligned} \eta_{cc0} &= P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) - \eta_{\bullet n} \\ &= P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) - P(Y_i = 1 \mid Z_i = 1, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0). \end{aligned}$$

Hence η_{cc0} (and similarly, η_{cc1}) are also identified under Assumptions 1.1 - 1.6 and 1.8. However, depending on which option we choose, Assumption 1.7 or 1.8, $ACE(cc)$ will be defined differently, as we will show in Section 1.4.

1.3 Validating Assumptions

The assumptions in Section 1.2 impose restrictions on the observed data distribution, which we will now explore. The graphical representation of the binary model under Assumptions 1.1 - 1.3 is shown in Figure 1.6. The nodes depict the observables, and the arrows depict all possible relationships between the observables. The model is depicted by a partially directed graph in order to purposely leave the direction of association between the provider's compliance (D_i^{P}) and subject's compliance (D_i^{S}) ambiguous. In other words, both parties in the provider-subject pair may influence the other's compliance. In the MI example, both the therapist and patient's compliance necessarily influences the other's responses, and hence, compliance.

The red, orange and green arrows in Figure 1.6 correspond to relationships that do not hold under particular assumptions in Section 1.2. Each of these relationships can be tested in the observed data. Assumption 1.6 implies Z_i only influences Y_i through D_i^{P} and D_i^{S}. In other words, Z_i is an *instrumental variable*. Similarly, Assumption 1.5 implies Z_i only influences D_i^{S} through D_i^{P} and Assumption 1.7 implies D_i^{P} only influences Y_i through D_i^{S}. Finally, Assumption 1.4 eliminates certain compliance types (which is not readily seen in the partial directed graph).

1.3.1 Validating Assumption 1.6

We first consider the relationship depicted by the red arrow that connects Z_i and Y_i in Figure 1.6. Assumption 1.6 implies Z is an instrumental variable, eliminating the red arrow in Figure 1.6. While Assumption 1.6 is not verifiable since it involves parameters of the unobserved data distribution, Pearl [1995] showed it is possible to test whether a model involving instrumental variables may account for the observed data. Specifically, while we cannot truely test that the Assumption 1.6 holds, we can evaluate if this assumption is consistent with the observed data. Pearl's proposed test checks whether a set of inequalities holds in the observed data, called the *instrumental inequalities*. However, Pearl only considered the cases

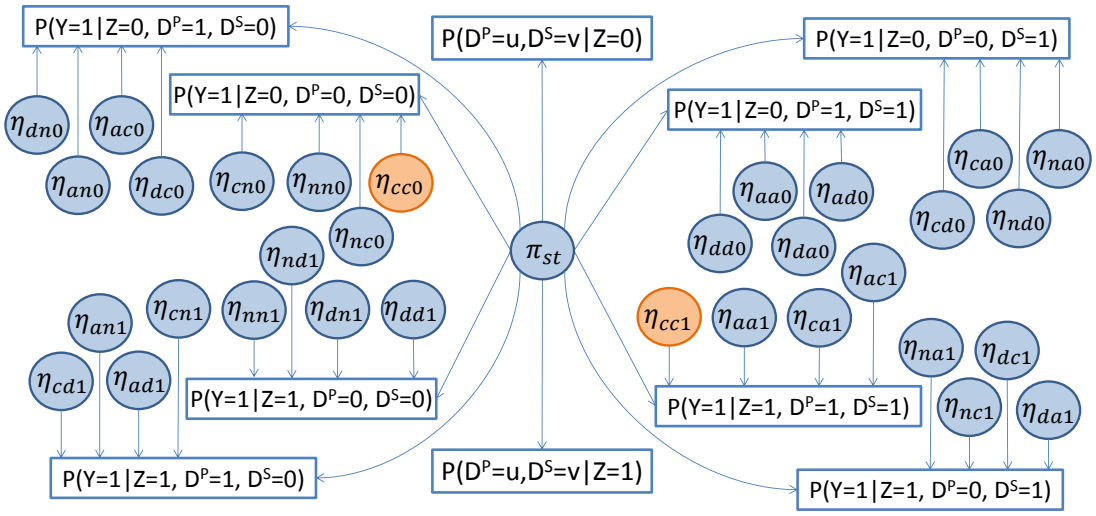


Figure 1.2: Illustration of the statistical model under Assumptions 1.1 - 1.3. Observed model parameters are represented by square nodes and the unobserved model parameters are represented by round nodes. Orange circular nodes correspond to the parameters that define $ACE(cc)$.

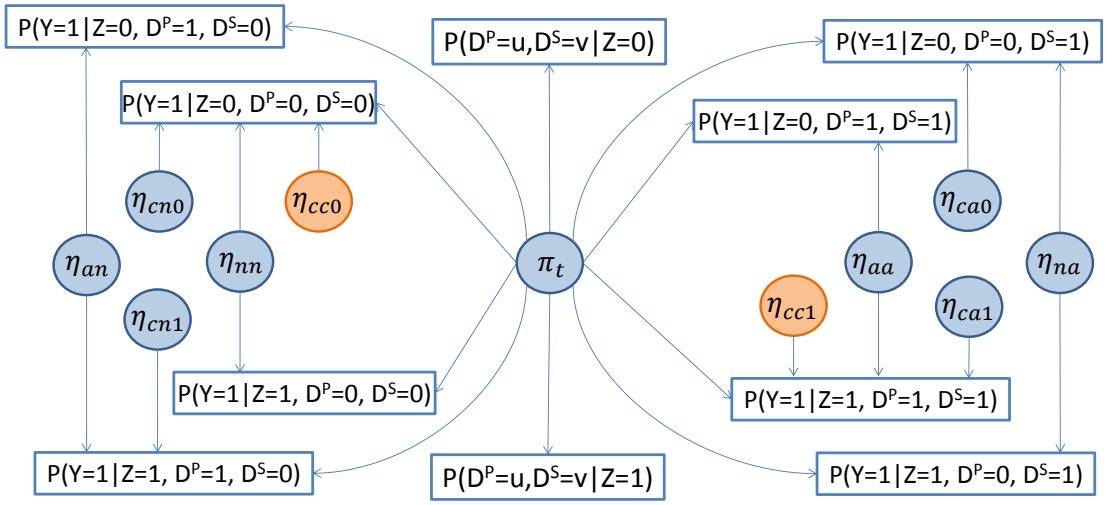


Figure 1.3: Illustration of the statistical model under Assumptions 1.1 - 1.6.

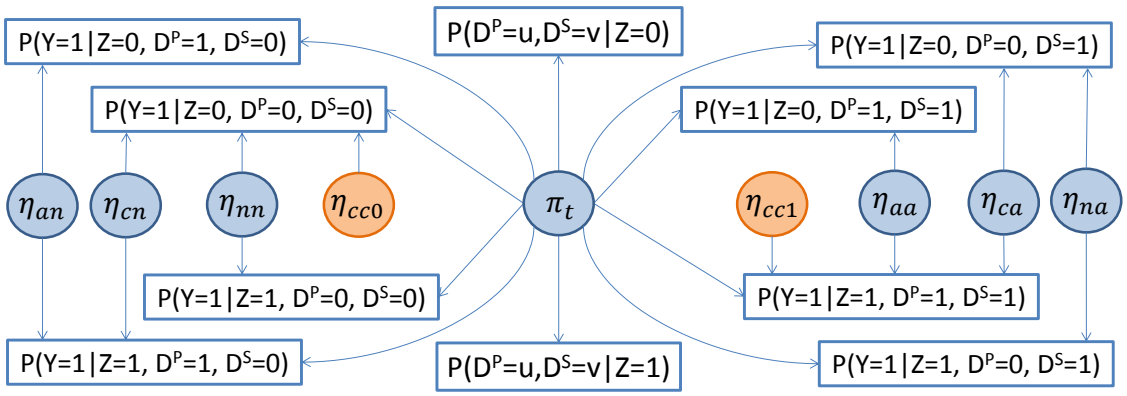


Figure 1.4: Illustration of the statistical model under Assumptions 1.1 - 1.6 & 1.7. Observed model parameters are represented by square nodes and the unobserved model parameters are represented by round nodes. Orange circular nodes correspond to the parameters that define $ACE(cc)$.

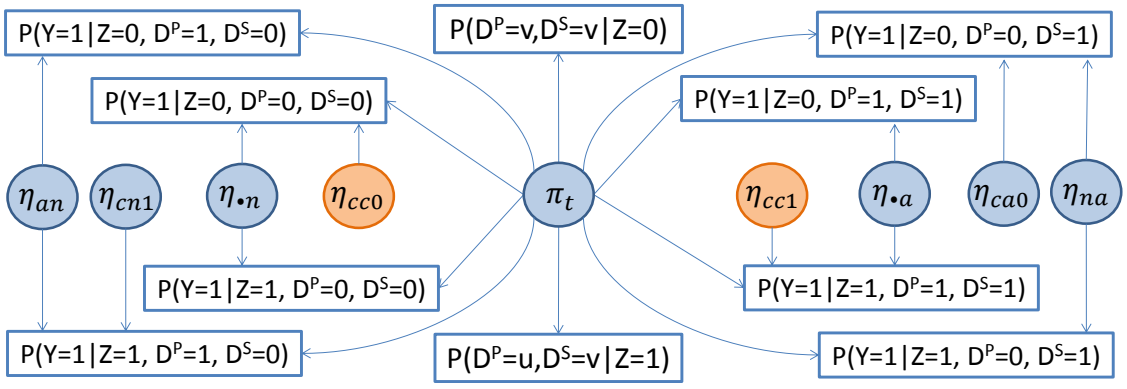


Figure 1.5: Illustration of the statistical model under Assumptions 1.1 - 1.6 & 1.8.

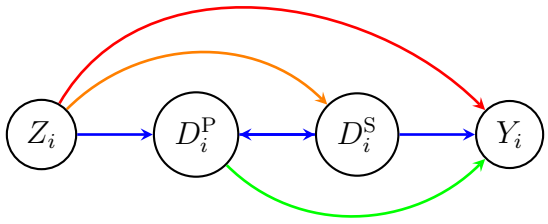


Figure 1.6: Partial directed graph that depicts the statistical model. The red, orange and green arrows indicate relationships that may exist if Assumptions 1.5, 1.6 and 1.7 do not hold, respectively.

with perfect provider compliance (i.e. $Z_i = D_i^{\mathrm{P}}$) or the case where the intermediate variable is binary. Bonet [2001] extended Pearl's test to settings where the intermediate variable takes more than two states, which includes the setting we are interested in (i.e. $(D_i^{\mathrm{P}}, D_i^{\mathrm{S}})$ takes one of four values). Based on the results of Bonet [2001], there are eight nontrivial instrumental inequalities that validate Assumption 1.6:

$$
\begin{aligned}
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=0)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=1) \leq 1\\
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=1)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=0) \leq 1\\
&\\
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1, Y_i=0 \mid Z_i=0)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1, Y_i=1 \mid Z_i=1) \leq 1\\
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1, Y_i=0 \mid Z_i=1)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1, Y_i=1 \mid Z_i=0) \leq 1\\
&\\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=0)+P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=1) \leq 1\\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=1)+P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=0) \leq 1\\
&\\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1, Y_i=0 \mid Z_i=0)+P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1, Y_i=1 \mid Z_i=1) \leq 1\\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1, Y_i=0 \mid Z_i=1)+P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1, Y_i=1 \mid Z_i=0) \leq 1.
\end{aligned}
\tag{1.2}
$$

To validate Assumption 1.6 in practice, the inequalities in (1.2) are checked with sample probabilities.

In the special case where treatment access is not possible when $z = 0$, that is $D_i^{\mathrm{P}} = D_i^{\mathrm{S}} = 0$ when $Z_i = 0$, only the first two inequalities in (1.2) are nontrivial. To see why, first note that when treatment access is not possible when $z = 0$, a natural coding of the potential compliance variables under $z = 0$ is $D_i^{\mathrm{P}}(0) = 0$ and $D_i^{\mathrm{S}}(0) = 0$ for all i. Hence $P(D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v \mid Z_i = 0) = 0$ for $u = 1$ and/or $v = 1$ and $P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) = 1$, making all but the first two inequalities trivial.

If a distribution fails to satisfy (1.2), it is not compatible with the model under Assumption 1.6. However, if the distribution does satisfy (1.2), it is possible that the data generating model makes other assumptions which lead to (1.2). Satisfying (1.2) does not necessarily imply Assumption 1.6 holds, but it does imply a weaker result: there is no evidence to suggest that the assumption does not hold.

1.3.2 Validating Assumption 1.5

Assumption 1.5 implies Z_i only influences D_i^{S} through D_i^{P}, eliminating the orange arrow that connects Z_i and D_i^{S} in Figure 1.6. Since each of these variables are binary, we can simply implement the instrumental inequalities of Pearl [1993] to validate Assumption 1.5:

$$\begin{aligned}
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0 \mid Z_i=0) + P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1 \mid Z_i=1) \leq 1 \\
&P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0 \mid Z_i=1) + P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1 \mid Z_i=0) \leq 1 \\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{P}}=0 \mid Z_i=0) + P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1 \mid Z_i=1) \leq 1 \\
&P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{P}}=0 \mid Z_i=1) + P(D_i^{\mathrm{P}}=1, D_i^{\mathrm{S}}=1 \mid Z_i=0) \leq 1.
\end{aligned} \tag{1.3}$$

In (1.3) we are assuming that treatment access is possible when $z=0$, that is, a provider may or may not follow treatment protocol and a subject may or may not take treatment. In the special case where treatment access is not possible when $z=0$, providers never follow treatment protocol and subjects never take treatment, that is $D_i^{\mathrm{P}} = D_i^{\mathrm{S}} = 0$ when $Z_i = 0$. In this case, the inequalities in (1.3) are trivial and the following equality holds:

$$P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1 | Z_i=1) = 0. \tag{1.4}$$

Denote potential compliance variables under $z=0$ by $D_i^{\mathrm{P}}(0)=0$ and $D_i^{\mathrm{S}}(0)=0$ for all i. Observing any provider-subject pair such that $Z_i=1$, $D_i^{\mathrm{P}}=0$ and $D_i^{\mathrm{S}}=1$ would imply the ith provider-subject pair is of type nc, which violates Assumption 1.5. The first inequality in (1.3) is equivalent to (1.4) and the last 3 inequalities are trivial in this setting since $P(D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v \mid Z_i = 0) = 0$ for $u=1$ and/or $v=1$ and $P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0 \mid Z_i=0) = 1$.

As in Section 1.3.1, if a distribution fails to satisfy the inequalities (1.3), it is not compatible with the model under Assumptions 1.5, however, satisfying the inequalities does not imply the assumption holds. Nevertheless, checking (1.3) constitutes an empirical test for violation of the model under Assumptions 1.5.

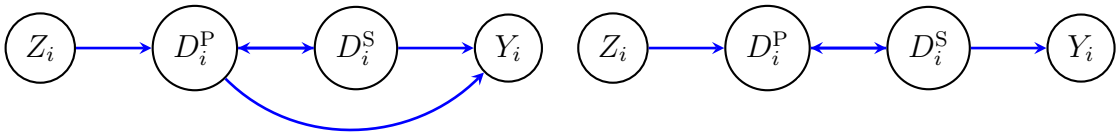
1.3.3 Validating Assumption 1.7

Assumption 1.7 implies D_i^{P} only influences Y_i through D_i^{S}, eliminating the green arrow connecting D_i^{P} to Y_i in Figure 1.6. Similar to Section 1.3.2, since each of these variables are binary, we can implement the instrumental inequalities of Pearl [1993] to test Assumption 1.7:

$$\begin{aligned}
&P(D_i^{\mathrm{S}}=0, Y_i=0 \mid D_i^{\mathrm{P}}=0) + P(D_i^{\mathrm{S}}=0, Y_i=1 \mid D_i^{\mathrm{P}}=1) \leq 1 \\
&P(D_i^{\mathrm{S}}=0, Y_i=0 \mid D_i^{\mathrm{P}}=1) + P(D_i^{\mathrm{S}}=0, Y_i=1 \mid D_i^{\mathrm{P}}=0) \leq 1 \\
&P(D_i^{\mathrm{S}}=1, Y_i=0 \mid D_i^{\mathrm{P}}=0) + P(D_i^{\mathrm{S}}=1, Y_i=1 \mid D_i^{\mathrm{P}}=1) \leq 1 \\
&P(D_i^{\mathrm{S}}=1, Y_i=0 \mid D_i^{\mathrm{P}}=1) + P(D_i^{\mathrm{S}}=1, Y_i=1 \mid D_i^{\mathrm{P}}=0) \leq 1.
\end{aligned} \tag{1.5}$$

The graphical representations of the model under Assumptions 1.1 - 1.6 and 1.7 vs. Assumptions 1.1 - 1.6 and 1.8 illustrates an important difference, shown in Figure 1.7. In the MI example, the extra edge connecting D_i^{P} and Y_i in Figure 1.7a compared no edge in Figure 1.7b indicates therapists' fidelity to the MI method may affect patient outcomes regardless of whether it affects the patient's expressed intentions during the therapy session. Contrarily, the missing edge in Figure 1.7b suggests that only the patient's expressed intentions during therapy matter, which is a strong assumption.

If a distribution fails to satisfy (1.5), it is not compatible with the model under Assumption 1.7. An example of such a distribution is one where $P(D_i^{\mathrm{S}}=0, Y_i=0 \mid D_i^{\mathrm{P}}=0) = P(D_i^{\mathrm{S}}=0, Y_i=1 \mid D_i^{\mathrm{P}}=1) = 0.6$. Satisfying the inequalities does not imply Assumption 1.7 holds. But, a violation of (1.5) suggests that 1.7 does not hold so Assumption 1.8 is at least not provably inconsistent with the observed data.



(a) Partial directed graph under Assumptions 1.1 - 1.6 & 1.8

(b) Partial directed graph under Assumptions 1.1 - 1.6 & 1.7

Figure 1.7: Partial directed graphs that depicts the statistical model under two different sets of assumptions considered here.

1.3.4 Validating Assumption 1.4

Like Assumptions 1.5, 1.6 and 1.7, Assumption 1.4 is not verifiable because it involves unobserved variables (i.e. $D^{\mathrm{P}}(z)$ and $D^{\mathrm{S}}(z)$ are only observed for either $z = 1$ or $z = 0$). Nevertheless Assumption 1.4 also has testable implications:

$$P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0|Z_i = 1) \geq P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0|Z_i = 0), \tag{1.6}$$

$$P(D_i^{\mathrm{S}} = 1|Z_i = 1) \geq P(D_i^{\mathrm{S}} = 1|Z_i = 0) \text{ or equivalently, } P(D_i^{\mathrm{S}} = 0|Z_i = 0) \geq P(D_i^{\mathrm{S}} = 0|Z_i = 1), \tag{1.7}$$

$$P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1|Z_i = 0) \geq P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1|Z_i = 1). \tag{1.8}$$

The first inequality (1.6) holds since the left and right hand sides are the proportions of types *cn* or *an* and just *an*, respectively, since assumptions 1.4 and 1.5 rule out any other possible types. Hence a violation of (1.6) would imply the existence of types *ad*, *cd*, *dn*, *dc* and/or *ac*. Similarly, the second inequality (1.7) holds since the left and right hand sides

are proportions of types $na/ca/aa/cc$ and $na/ca/aa$ (or equivalently, proportions of types $nn/cn/an/cc$ and $nn/cn/an$), respectively. Lastly, the third inequality (1.8) holds since the left and right hand sides are proportions of types na/ca and na. The inequalities (1.6), (1.7) and (1.8) could be considered an extension to those that hold in the perfect provider compliance setting [Angrist and Imbens, 1995, Richardson et al., 2011].

The inequalities (1.6), (1.7) and (1.8) have an intuitive relationship with Assumption 1.4. The inequality (1.7) is what one would expect under Assumption 1.4. Furthermore, inequalities (1.7) and (1.8) imply $P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1|Z_i = 1) \geq P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1|Z_i = 0)$. Together with (1.6), this implies $P(D_i^{\mathrm{P}} = 1|Z_i = 1) \geq P(D_i^{\mathrm{P}} = 1|Z_i = 0)$, which one would also expect under Assumption 1.4. In the case where treatment access is not possible when $z = 0$, all three of the inequalities are trivial.

Similar to the previous tests, if a distribution fails to satisfy the inequalities (1.6) - (1.8), it is not compatible with the model under Assumptions 1.4. However, once again, satisfying the inequalities does not imply the Assumption 1.4 holds.

1.4 Proposed Estimators for $ACE(cc)$

We consider estimation of $ACE(cc)$ under two cases: (1) settings where provider-subject pairs may access treatment when $z = 0$, and (2) settings where there is no access to treatment when $z = 0$. For each case, we develop MOM and ML estimators.

1.4.1 Treatment Access is Possible when $z = 0$

In some settings, it is possible for providers/subjects to follow protocol/take treatment when assigned to $z = 0$. For example, suppose the mechanism for randomization to treatment is randomization of therapist-patient pairs to a therapist training in a new method. Some therapists may naturally implement the method being trained and some patients may participate in therapy regardless of the therapist's methods. In this case, the MOM estimator under Assumptions 1.1 - 1.6 and 1.7 was proposed in Schochet and Chiang [2011]. In this section, we will propose both a MOM and ML estimator under Assumptions 1.1 - 1.6 and

1.8.

Proposed MOM Estimator for $ACE(cc)$

The MOM estimator under Assumptions 1.1 - 1.6 and 1.8 is a consequence of the proof of Proposition 1.2. Denote the observed data by $\mathbf{n}$, a vector indexed by $z, u, v, y \in \{0, 1\}$ where n_{zuvy} denotes the count of observed data quadruples ($Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v$, $Y_i = y$). Since each observable is binary, there are $2^4 = 16$ possible combinations. Then $\mathbf{n} \sim Multinomial_{16}(n, \boldsymbol{\rho})$ where $\boldsymbol{\rho}$ is the 16-vector indexed by $z, u, v, y \in \{0, 1\}$ where $\rho_{zuvy} \equiv P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y)$. The 8 cells corresponding to $Y_i = 1$ are depicted as the 4 left-most and 4 right-most rectangular nodes in Figure 1.5 (the 8 cells corresponding to $Y_i = 0$ are not shown in the figure). By Proposition 1.2, $ACE(cc) = \eta_{cc1} - \eta_{cc0}$ can be written as a function of $\boldsymbol{\rho}$. By the law of large numbers and Slutsky's theorem, asymptotically consistent estimators of η_{cc0} and η_{cc1} are:

$$\begin{aligned}
\eta_{cc0}^{\mathrm{MOM}} &= \frac{(n_{0001}n_{100\bullet} + n_{1001}n_{010\bullet})n_{1\cdots} - (n_{1001}n_{100\bullet} + n_{1001}n_{110\bullet})n_{0\cdots}}{(n_{000\bullet}n_{100\bullet} + n_{100\bullet}n_{010\bullet})n_{1\cdots} - (n_{100\bullet}n_{100\bullet} + n_{100\bullet}n_{110\bullet})n_{0\cdots}} \\
\eta_{cc1}^{\mathrm{MOM}} &= \frac{(n_{1111}n_{011\bullet} + n_{0111}n_{101\bullet})n_{0\cdots} - (n_{0111}n_{011\bullet} + n_{0111}n_{001\bullet})n_{1\cdots}}{(n_{111\bullet}n_{011\bullet} + n_{011\bullet}n_{101\bullet})n_{0\cdots} - (n_{011\bullet}n_{011\bullet} + n_{011\bullet}n_{001\bullet})n_{1\cdots}},
\end{aligned}$$

where $n_{zuv\bullet} = \sum_{y=0,1} n_{zuvy}$ and $n_{z\cdots} = \sum_{u=0,1}\sum_{v=0,1}\sum_{y=0,1} n_{zuvy}$. The MOM estimator for $ACE(cc) = \eta_{cc1} - \eta_{cc0}$ is then

$$ACE(cc)_{\mathrm{MOM}} = \eta_{cc1}^{\mathrm{MOM}} - \eta_{cc0}^{\mathrm{MOM}}.$$

An estimator for the variance of $ACE(cc)_{\mathrm{MOM}}$ can be obtained via the multivariate central limit theorem and delta method. The covariance matrix of $\mathbf{n}$ is $n(\mathrm{diag}(\boldsymbol{\rho}) - \boldsymbol{\rho}\boldsymbol{\rho}')$. By the multivariate central limit theorem,

$$\sqrt{n}(\mathbf{n}/n - \boldsymbol{\rho}) \to_d \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\rho}) - \boldsymbol{\rho}\boldsymbol{\rho}').$$

Hence by the delta method, an asymptotically correct estimator for the variance of $ACE(cc)_{\mathrm{MOM}}$ is

$$\widehat{Var}(ACE(cc)_{\mathrm{MOM}}) = \boldsymbol{\Delta}(\mathbf{n}/n)\left(\frac{1}{n}\mathrm{diag}(\mathbf{n}) - \frac{1}{n^2}\mathbf{n}\mathbf{n}'\right)\boldsymbol{\Delta}(\mathbf{n}/n)',$$

where $\boldsymbol{\Delta}(\mathbf{n}/n)$ is a vector indexed by $z, u, v, y \in \{0,1\}$ with $\Delta_{zuvy}(\mathbf{n}/n) \equiv \frac{\partial ACE(cc)}{\partial \rho_{zuvy}}\big|_{\boldsymbol{\rho}=\mathbf{n}/n}$. As evident from the formulas for η_{cc0}^{MOM} and η_{cc1}^{MOM}, the closed form expression of $\boldsymbol{\Delta}(\mathbf{n}/n)$ is lengthy and so we do not reproduce it here. The expression can easily be obtained via a computer algebra system (e.g. in R we utilized YACAS via the `Ryacas` package).

Proposed ML Estimator for $ACE(cc)$

By considering compliance type as a missing covariate, the EM-algorithm [Dempster et al., 1977] is well suited to ML estimation of $ACE(cc)$. Under Assumptions 1.1 - 1.6 and 1.8, the parameters of the model are $\theta = (\xi, \eta, \pi)$ where: $\xi = P(Z_i = 1)$; η is a vector of expected outcomes conditional on compliance type and treatment assignment, $\eta = (\eta_{cc0}, \eta_{cc1}, \eta_{cn1}, \eta_{ca0}, \eta_{an}, \eta_{na}, \eta_{\bullet n}, \eta_{\bullet a})$; π is a vector of probabilities of each compliance type, $\pi = (\pi_{cc}, \pi_{cn}, \pi_{ca}, \pi_{an}, \pi_{na}, \pi_{nn}, \pi_{aa})$. The complete data likelihood, L, is defined as follows:

$$
\begin{aligned}
L \;=\; & \prod_{i=1}^{n} (1-\xi)^{\mathbb{1}(Z_i=0)} \times \xi^{\mathbb{1}(Z_i=1)} \times \\
& \left([\eta_{\bullet n}^{Y_i}(1-\eta_{\bullet n})^{1-Y_i}]^{\mathbb{1}(C_i\in\{nn,cn\})}\pi_{nn}^{\mathbb{1}(C_i=nn)}\pi_{cn}^{\mathbb{1}(C_i=cn)}\right)^{\mathbb{1}(Z_i=0,D_i^{\text{P}}=0,D_i^{\text{S}}=0)} \times \\
& \left(\eta_{\bullet n}^{Y_i}(1-\eta_{\bullet n})^{1-Y_i}\pi_{nn}\right)^{\mathbb{1}(C_i=nn,Z_i=1,D_i^{\text{P}}=0,D_i^{\text{S}}=0)} \left(\eta_{an}^{Y_i}(1-\eta_{an})^{1-Y_i}\pi_{an}\right)^{\mathbb{1}(C_i=an,Z_i=0,D_i^{\text{P}}=1,D_i^{\text{S}}=0)} \times \\
& \left([\eta_{na}^{Y_i}(1-\eta_{na})^{1-Y_i}\pi_{na}]^{\mathbb{1}(C_i=na)}[\eta_{ca0}^{Y_i}(1-\eta_{ca0})^{1-Y_i}\pi_{ca}]^{\mathbb{1}(C_i=ca)}\right)^{\mathbb{1}(Z_i=0,D_i^{\text{P}}=0,D_i^{\text{S}}=1)} \times \\
& \left([\eta_{an}^{Y_i}(1-\eta_{an})^{1-Y_i}\pi_{an}]^{\mathbb{1}(C_i=an)}[\eta_{cn1}^{Y_i}(1-\eta_{cn1})^{1-Y_i}\pi_{cn}]^{\mathbb{1}(C_i=cn)}\right)^{\mathbb{1}(Z_i=1,D_i^{\text{P}}=1,D_i^{\text{S}}=0)} \times \\
& \left(\eta_{\bullet a}^{Y_i}(1-\eta_{\bullet a})^{1-Y_i}\pi_{aa}\right)^{\mathbb{1}(C_i=aa,Z_i=0,D_i^{\text{P}}=1,D_i^{\text{S}}=1)} \left(\eta_{na}^{Y_i}(1-\eta_{na})^{1-Y_i}\pi_{na}\right)^{\mathbb{1}(C_i=na,Z_i=1,D_i^{\text{P}}=0,D_i^{\text{S}}=1)} \times \\
& \left([\eta_{\bullet a}^{Y_i}(1-\eta_{\bullet a})^{1-Y_i}]^{\mathbb{1}(C_i\in\{aa,ca\})}\pi_{aa}^{\mathbb{1}(C_i=aa)}\pi_{ca}^{\mathbb{1}(C_i=ca)}\right)^{\mathbb{1}(Z_i=1,D_i^{\text{P}}=1,D_i^{\text{S}}=1)} \times \\
& \left(\eta_{cc0}^{Y_i}(1-\eta_{cc0})^{1-Y_i}\pi_{cc}\right)^{\mathbb{1}(C_i=cc,Z_i=0,D_i^{\text{P}}=0,D_i^{\text{S}}=0)} \left(\eta_{cc1}^{Y_i}(1-\eta_{cc1})^{1-Y_i}\pi_{cc}\right)^{\mathbb{1}(C_i=cc,Z_i=1,D_i^{\text{P}}=1,D_i^{\text{S}}=1)}.
\end{aligned}
$$

In the kth E-step, the conditional expectation of the complete data log-likelihood given the observed data under the current parameter value, $\theta^{(k)}$, is computed. In the M-step, the update $\theta^{(k+1)}$ maximizes the computed expectation. The E- and M-steps are repeated until convergence. Multiple starting values for $\theta^{(0)}$ are tried, the final update that corresponds to the highest converged likelihood value is considered the ML estimate that corresponds

to the global maximum, which we denote by $\hat{\theta}$. (Details of the EM algorithm are given in Appendix A).

An asymptotically correct estimate of the variance-covariance matrix of $\hat{\theta}$ can be estimated from the inverse of the observed data information matrix, $I_{\mathbf{n}}$. Since computing the gradient or Hessian of the observed data likelihood is cumbersome, we use Louis's formula for Multinomial data [Louis, 1982] to obtain $I_{\mathbf{n}}$. The ML estimate, denoted by $ACE(cc)_{\text{ML}}$ is thus

$$ACE(cc)_{\text{ML}} = \hat{\eta}_{cc1} - \hat{\eta}_{cc0},$$

and an asymptotically consistent estimate of its variance is

$$\widehat{Var}(ACE(cc)_{\text{ML}}) = (1,-1)I_{\mathbf{n}}^{-1}[\hat{\eta}_{cc1}, \hat{\eta}_{cc0}](1,-1)^T,$$

where $I_{\mathbf{n}}^{-1}[\hat{\eta}_{cc1}, \hat{\eta}_{cc0}]$ is the 2×2 submatrix of $I_{\mathbf{n}}^{-1}$ with rows/columns that correspond to the parameters $\hat{\eta}_{cc1}$ and $\hat{\eta}_{cc0}$. (Details for obtaining $I_{\mathbf{n}}$ and $I_{\mathbf{n}}^{-1}[\hat{\eta}_{cc1}, \hat{\eta}_{cc0}]$ are given in Appendix A.)

1.4.2 Treatment Access is Not Possible when $z = 0$

In some settings, it is not possible for providers/subjects to access treatment when assigned to $z = 0$. For example, consider a study where therapist-patient pairs are randomized to therapist training and a behavioral intervention vs. no intervention. In this setting, no therapy intervention occurs when $z = 0$, and hence it is reasonable to assume $(D_i^P, D_i^S) = (0, 0)$ for i such that $Z_i = 0$. The number of compliance types under Assumptions Assumptions 1.1 - 1.6 and 1.8 is thus reduced to three types: nn, cn and cc.

Proposed MOM Estimator for $ACE(cc)$

By Proposition 1.2, the law of large numbers and Slutsky's theorem, it can be shown that an asymptotically consistent MOM estimator under Assumptions 1.1 - 1.6 and 1.8 is:

$$ACE(cc)_{\text{MOM}} = \frac{n_{1111}}{n_{111\bullet}} - \frac{n_{0001}n_{100\bullet}n_{1\bullet\bullet\bullet} - n_{1001}n_{0\bullet\bullet\bullet}(n_{110\bullet} - n_{100\bullet})}{n_{000\bullet}n_{100\bullet}n_{1\bullet\bullet\bullet} - n_{100\bullet}n_{0\bullet\bullet\bullet}(n_{110\bullet} - n_{100\bullet})},$$

while the natural extension of the proposal of Schochet and Chiang [2011] to the setting where providers and subjects in the untreated arm cannot crossover to treatment is:

$$ACE(cc)^{\text{SC}}_{\text{MOM}} = \frac{n_{1111}}{n_{111\bullet}} - \frac{n_{1\bullet\bullet\bullet}n_{0001} - n_{0\bullet\bullet\bullet}(n_{1101} + n_{1001})}{n_{1\bullet\bullet\bullet}n_{000\bullet} - n_{0\bullet\bullet\bullet}(n_{110\bullet} + n_{100\bullet})},$$

where $n_{zuv\bullet} = \sum_{y=0,1} n_{zuvy}$ and $n_{z\bullet\bullet\bullet} = \sum_{u=0,1}\sum_{v=0,1}\sum_{y=0,1} n_{zuvy}$. An estimator for the variance of $ACE(cc)_{\text{MOM}}$ or $ACE(cc)^{\text{SC}}_{\text{MOM}}$ can be obtained via the multivariate central limit theorem and delta method, similar to Section 1.4.1.

Proposed ML Estimator for $ACE(cc)$

Similar to Section 1.4.1, ML estimates of $ACE(cc)$ can be obtained from the EM algorithm. (In fact, the EM algorithm procedure is similar regardless of assuming Assumptions 1.1 - 1.6 and 1.7 or 1.8.) Under Assumptions 1.1 - 1.6 and 1.8, the complete data likelihood is:

$$\begin{aligned} L \;=\; & \prod_{i=1}^{n}(1-\xi)^{\mathbb{1}(Z_i=0)} \times \xi^{\mathbb{1}(Z_i=1)} \times \\ & \left([\eta_{\bullet n}^{Y_i}(1-\eta_{\bullet n})^{1-Y_i}]^{\mathbb{1}(C_i\in\{nn,cn\})}\pi_{nn}^{\mathbb{1}(C_i=nn)}\pi_{cn}^{\mathbb{1}(C_i=cn)}\right)^{\mathbb{1}(Z_i=0,D_i^{\text{P}}=0,D_i^{\text{S}}=0)} \times \\ & \left(\eta_{\bullet n}^{Y_i}(1-\eta_{\bullet n})^{1-Y_i}\pi_{nn}\right)^{\mathbb{1}(C_i=nn,Z_i=1,D_i^{\text{P}}=0,D_i^{\text{S}}=0)}\left(\eta_{cn1}^{Y_i}(1-\eta_{cn1})^{1-Y_i}\pi_{cn}\right)^{\mathbb{1}(C_i=cn,Z_i=1,D_i^{\text{P}}=1,D_i^{\text{S}}=0)} \times \\ & \left(\eta_{cc0}^{Y_i}(1-\eta_{cc0})^{1-Y_i}\pi_{cc}\right)^{\mathbb{1}(C_i=cc,Z_i=0,D_i^{\text{P}}=0,D_i^{\text{S}}=0)}\left(\eta_{cc1}^{Y_i}(1-\eta_{cc1})^{1-Y_i}\pi_{cc}\right)^{\mathbb{1}(C_i=cc,Z_i=1,D_i^{\text{P}}=1,D_i^{\text{S}}=1)}. \end{aligned}$$

The ML estimate of $ACE(cc)$ as well as an asymptotically correct estimate of the variance of $ACE(cc)_{\text{ML}}$ can be obtained via the EM algorithm and Louis's formula, similar to Section 1.4.1.

1.5 Simulation Studies

We performed simulation studies to assess the bias and efficiency of $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ compared to the estimator proposed by Schochet and Chiang [2011], denoted by $ACE(cc)^{\text{SC}}_{\text{MOM}}$. We consider six scenarios. In the first three, treatment access is possible in the $z = 0$ arm of the RCT. In the last three, treatment access is not possible when $z = 0$. In each scenario, we simulated data from a Multinomial distribution with parameter values

chosen so that Assumptions 1.1 - 1.6 hold. In scenarios 1 and 4, both Assumptions 1.7 and 1.8 hold. In scenarios 2 and 5, Assumption 1.8 holds but Assumption 1.7 does not hold. In scenarios 3 and 6, Assumption 1.7 holds but Assumption 1.8 does not hold. Hence the assumptions of $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ are met in scenarios 1, 2, 4 and 5, while the assumptions of $ACE(cc)^{\text{SC}}_{\text{MOM}}$ are met in scenarios 1, 3, 4 and 6.

For each scenario, we simulated 1000 complete datasets from sample sizes ranging from $n = 100$ to $n = 100,000$. Observed data $(Z_i, D_i^{\text{P}}, D_i^{\text{S}}, Y_i)$ were constructed from the complete data (Z_i, C_i, Y_i). Each of the three estimators, $ACE(cc)_{\text{MOM}}$, $ACE(cc)^{\text{SC}}_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ were evaluated on each observed dataset to obtain estimates of $ACE(cc)$ and estimates of the estimators' variance. The parameters of the data generating model are π, η, ξ, where $|\pi| = 7, |\eta| = 14$ in scenarios 1-3 and $|\pi| = 3, |\eta| = 6$ in scenarios 4-6. Each dataset was comprised of n independent draws from a $Multinomial_{28}(n, \boldsymbol{\rho}^*)$ in scenarios 1-3 or $Multinomial_{12}(n, \boldsymbol{\rho}^*)$ in scenarios 4-6. In each Multinomial distribution, $\boldsymbol{\rho}^*$ is a probability vector indexed by st, z, and y with $\rho^*_{stzy} \equiv \xi^z(1-\xi)^{1-z}\eta^y_{stz}(1-\eta_{stz})^{1-y}\pi_{st}$. We set the value of $\xi = 0.5$, $\pi_{st} = 1/|\pi|$ for all π_{st}, and the value of η varies by scenario, shown in Table 1.1.

The bias, MSE and coverage of 95% confidence intervals based on the 2.5% and 97.5% quantiles of a normal distribution with the estimated variance were computed. Results are shown in Tables 1.2 and 1.3. The simulation results suggest that when Assumptions 1.1 - 1.8 hold, all three estimators approach the true $ACE(cc)$ and 95% confidence intervals achieve approximately 95% coverage. Negligible bias, MSE and appropriate 95% confidence interval coverage occurred with finite sample sizes $n = 500$ and $n = 100$ when treatment access was and was not possible when $z = 0$, respectively. However, when Assumption 1.7 did not hold, $ACE(cc)^{\text{SC}}_{\text{MOM}}$ was biased and coverage was poor. Similarly, when Assumption 1.8 did not hold, $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ were biased and coverage was poor.

Scenario	parameter values
1	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.3$ $\eta_{nn0} = 0.3$ $\eta_{ca0} = 0.5$ $\eta_{an0} = 0.3$ $\eta_{na0} = 0.5$ $\eta_{aa0} = 0.5$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.3$ $\eta_{nn1} = 0.3$ $\eta_{ca1} = 0.5$ $\eta_{an1} = 0.3$ $\eta_{na1} = 0.5$ $\eta_{aa1} = 0.5$
2	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.05$ $\eta_{nn0} = 0.05$ $\eta_{ca0} = 0.5$ $\eta_{an0} = 0.2$ $\eta_{na0} = 0.4$ $\eta_{aa0} = 0.6$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.1$ $\eta_{nn1} = 0.05$ $\eta_{ca1} = 0.6$ $\eta_{an1} = 0.2$ $\eta_{na1} = 0.4$ $\eta_{aa1} = 0.6$
3	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.1$ $\eta_{nn0} = 0.05$ $\eta_{ca0} = 0.53$ $\eta_{an0} = 0.2$ $\eta_{na0} = 0.56$ $\eta_{aa0} = 0.6$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.1$ $\eta_{nn1} = 0.05$ $\eta_{ca1} = 0.53$ $\eta_{an1} = 0.2$ $\eta_{na1} = 0.56$ $\eta_{aa1} = 0.6$
4	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.3$ $\eta_{nn0} = 0.3$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.3$ $\eta_{nn1} = 0.3$
5	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.05$ $\eta_{nn0} = 0.05$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.1$ $\eta_{nn1} = 0.05$
6	$\eta_{cc0} = 0.3$ $\eta_{cn0} = 0.1$ $\eta_{nn0} = 0.05$ $\eta_{cc1} = 0.5$ $\eta_{cn1} = 0.1$ $\eta_{nn1} = 0.05$

Table 1.1: Values of η used in each of six scenarios considered in the simulation studies of Section 1.5. In scenarios 1-3, $\eta = (\eta_{stz} : st \in \mathcal{C}_7 \text{ and } z \in \{0,1\})$. In scenarios 4-6, $\eta = (\eta_{stz} : st \in \{cc, cn, nn\} \text{ and } z \in \{0,1\}$.

1.6 Application to a RCT of MI

As a specific example, we applied our estimators of $ACE(cc)$ to data from a RCT of MI, where treatment access was not possible when $z = 0$ [Krupski et al., 2012]. The primary aim of the study was to determine efficacy of an MI-based intervention among patients of low socio-economic status presenting to primary care. A total of 868 individuals attending a scheduled medical care appointment that met problem drug use criteria were recruited. 435 patients were randomized to receive an intervention, a 30-minute MI-based counseling session and follow-up phone booster delivered by a trained counselor in addition to the usual care provided by their physician. 433 patients were randomized to control, the usual care provided by their physician. Since coding a MI session is costly, only 57 randomly selected sessions among 435 had nonmissing compliance and outcome data. 389 of the 433 in the control condition had nonmissing data since compliance data is not collected (automatically coded as $D_i^{\mathrm{P}} = D_i^{\mathrm{S}} = 0$ for all pairs) but some follow up data was missing.

Patients were randomized to receive MI (Z_i=1 if the patient from the ith therapist-patient pair was assigned to the intervention, $Z_i = 0$ otherwise). Provider compliance to the treatment protocol was measured according to five commonly used performance benchmarks based on the MI Skills Code (MISC) of the MI intervention. The MISC consists of global scales, e.g. a rating of the therapist's empathy, and counts of specific behaviors. To compute the latter, the transcript of the MI intervention is parsed into utterances, or complete thoughts. The five performance measures are: (1) an empathy rating from 0-5, (2) the ratio of reflective statements to questions asked, (3) the proportion of utterances that are open questions, (4) the proportion of utterances that are complex reflections, and (5) the proportion of utterances that are MI-consistent, as defined in the MISC manual. We define $D_i^{\mathrm{P}} = 1$ if the therapist met the majority of the suggested benchmarks for proficiency (at least three of the five) and $D_i^{\mathrm{P}} = 0$ otherwise or if $Z_i = 0$. Patient compliance was measured by the patient's expressed intentions during the counseling session. For each intervention, the number of utterances that expressed intent to change ('change talk') and the number

	$ACE(cc)_{\text{MOM}}$			$ACE(cc)^{\text{SC}}_{\text{MOM}}$			$ACE(cc)_{\text{ML}}$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
Scenario 1: Assumptions 1.1 - 1.8 hold									
500	0.021	0.37	0.991	0.040	0.48	0.984	-0.0091	0.11	0.992
2000	-0.0090	0.025	0.963	-0.0036	0.025	0.959	-0.0092	0.025	0.963
20000	0.00052	0.0022	0.956	-0.00023	0.0022	0.958	0.00052	0.0022	0.956
Scenario 2: Assumptions 1.1 - 1.6 and 1.8 hold, but Assumption 1.7 does not hold									
500	-0.040	0.20	0.967	0.14	0.17	0.949	-0.027	0.083	0.973
2000	0.0036	0.020	0.949	0.16	0.045	0.766	0.0034	0.020	0.950
20000	-0.00070	0.0017	0.959	0.15	0.024	0.058	-0.00068	0.0017	0.959
Scenario 3: Assumptions 1.1 - 1.6 and 1.7 hold, but Assumption 1.8 does not hold									
500	-0.19	0.26	0.984	-0.033	0.17	0.972	-0.15	0.12	0.985
2000	-0.14	0.041	0.918	-0.0056	0.020	0.958	-0.14	0.041	0.917
20000	-0.12	0.017	0.208	-0.0014	0.0019	0.953	-0.12	0.017	0.208

Table 1.2: Simulation results under three scenarios when treatment access is possible when $z = 0$. In Scenario 1, Assumptions 1.1 - 1.8 hold. In Scenario 2, Assumptions 1.1 - 1.6 and 1.8 hold, but Assumption 1.7 does not hold. In Scenario 3, Assumptions 1.1 - 1.6 and 1.7 hold, but Assumption 1.8 does not hold.

	$ACE(cc)_{\text{MOM}}$			$ACE(cc)^{SC}_{\text{MOM}}$			$ACE(cc)_{\text{ML}}$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
Scenario 4: Assumptions 1.1 - 1.8 hold									
100	-0.0011	0.13	0.953	-0.00082	0.098	0.958	-0.031	0.091	0.979
1000	-0.00023	0.010	0.949	-0.0015	0.0078	0.952	-0.00023	0.010	0.949
10000	-0.00074	0.0011	0.941	-0.00060	0.00083	0.943	-0.00074	0.0011	0.941
Scenario 5: Assumptions 1.1 - 1.6 and 1.8 hold, but Assumption 1.7 does not hold									
100	-0.0054	0.095	0.951	0.022	0.078	0.944	-0.023	0.073	0.978
1000	-0.0033	0.0075	0.953	0.022	0.0071	0.924	-0.0033	0.0075	0.953
10000	-0.00039	0.00081	0.942	0.025	0.0019	0.648	-0.00038	0.00081	0.942
Scenario 6: Assumptions 1.1 - 1.6 and 1.7 hold, but Assumption 1.8 does not hold									
100	-0.064	0.068	0.949	-0.0086	0.058	0.945	-0.066	0.063	0.982
1000	-0.052	0.0081	0.894	-0.0017	0.0049	0.945	-0.052	0.0081	0.894
10000	-0.050	0.003	0.419	-0.00047	0.00049	0.940	-0.05	0.003	0.419

Table 1.3: Simulation results under three scenarios when treatment access is not possible when $z = 0$. In Scenario 4, Assumptions 1.1 - 1.8 hold. In Scenario 5, Assumptions 1.1 - 1.6 and 1.8 hold, but Assumption 1.7 does not hold. In Scenario 6, Assumptions 1.1 - 1.6 and 1.7 hold, but Assumption 1.8 does not hold.

of utterances that expressed resistance to change ('sustain talk') were recorded. We defined $D_i^{\mathrm{S}} = 1$ if the ratio of change to sustain talk statements exceeded 2 since at least twice as much change talk compared to sustain talk is a clinically intuitive indicator that the patient is motivated to reduce their drug use; $D_i^{\mathrm{S}} = 0$ otherwise or if $Z_i = 0$. The primary outcome of interest was patients' self-reported days of drug use in the past 30 days, which was measured at randomization, and at 3, 6, 9 and 12-months from randomization. We define Y_i as an indicator of a decrease of at least 50% in days of drug use out of the last 30 days, comparing 3-month follow up to baseline. A decrease of at least 50% is considered clinically meaningful and the most recent follow-up time was used because the intervention is presumed to have a greater impact sooner rather than later.

The primary study aim is to estimate the causal effect of MI on patient outcomes. If a provider is unable to meet minimal proficiency standards regardless of training, or a patient is unwilling to consider changing the their drug use no matter how skilled the provider, it is unlikely that MI will have any impact on the patient outcome and an ITT analysis of the data will be biased toward a null intervention. The ITT analysis finds $ACE(cc) = 0.02$ (95% CI:[-0.16,0.12]), where 40.0% of subjects reduced drug use when assigned to MI counseling compared to 38.0% of subjects that were not assigned to counseling. However, provider and subject noncompliance were clearly problematic; in the $z = 1$ arm of the study, only 27.3% of patients became motivated to change and 81.8% of therapists met the majority of the five MI proficiency standards.

To address the two-level noncompliance in this example, we assume Assumptions 1.1 - 1.6 and argue Assumption 1.8 is more appropriate than Assumption 1.7. For a therapist-patient pair of type cn, although the patient is never motivated to change regardless of whether $z = 0$ or $z = 1$, it is likely that an intervention with a MI-adherent therapist would have a more positive effect on the patient's future drug use compared to an intervention with a non-adherent therapist or no intervention at all. Hence it is more plausible that Assumption 1.8 holds than Assumption 1.7 and our prosed methods of estimating the $ACE(cc)$ are more appropriate than the proposal of Schochet and Chiang [2011].

Under Assumptions 1.1 - 1.6 and 1.8, the first two inequalities in (1.2) and the equality (1.4) must hold, which we find is the case:

$$P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=0)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=1)=\frac{241}{389}+\frac{4}{57}=0.69\leq 1$$

$$P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=0 \mid Z_i=1)+P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=0, Y_i=1 \mid Z_i=0)=\frac{6}{57}+\frac{148}{389}=0.49\leq 1$$

$$P(D_i^{\mathrm{P}}=0, D_i^{\mathrm{S}}=1 \mid Z_i=1)=\frac{0}{57}.$$

However, note that two two provider-subject pairs were observed to have $D_i^{\mathrm{P}}=0$, $D_i^{\mathrm{S}}=1$ and $Z_i=1$ but barely exceeded the threshold for making $D_i^{\mathrm{S}}=1$. We attribute this to rounding error rather than a violation of (1.4) and Assumption 1.5. For these two pairs we have recoded $D_i^{\mathrm{S}}=0$.

In this example the $ACE(cc)_{\mathrm{MOM}}$ and $ACE(cc)_{\mathrm{ML}}$ were similar, both estimate the average causal effect of MI to be 0.14 (95% CI: [-0.73,1]). Moreover, the proportion of therapist-patient pairs of type *cc* was estimated to be 27.0% and among these pairs, 46.7% of subjects reduced drug use with proficient MI counseling compared to 32.8% of subjects with non-proficient or no counseling. Note that there were only 17 therapists in the study, so the therapist-patient pairs may not have been iid because patients were nested within therapists. Hence, whether Assumption 1.3 holds may be questionable, and as a result, the 95% confidence interval may be anticonservative. As an alternative, we computed the bootstrapped CIs based on 1000 bootstrap resamples. Each resample was a random sample with replacement of the 446 complete cases of the data, also of size 446. The 2.5 to 97.5 percentile of the $ACE(cc)_{\mathrm{MOM}}$ estimates computed on the bootstrap resamples ranged [-0.86,1.00], while the 2.5 to 97.5 percentile of the $ACE(cc)_{\mathrm{ML}}$ estimates were between [-0.64,0.64]. The conclusion based on the bootstrapped CIs is the same as the conclusion based on the above reported CI estimate: the estimated causal effect of MI is not statistically significant.

1.7 Final Remarks

In this chapter, we proposed new statistical methodology for making causal inference in RCTs where provider and subject compliance underlie the definition of treatment receipt, yet

provider and/or subject noncompliance is problematic. $ACE(cc)$, which represents the causal effect of treatment on provider-subject pairs that comply with assignment, can be identified under the assumptions outlined in Schochet and Chiang [2011], however these assumptions may not hold in some applications, specifically cognitive behavioral interventions. As a motivating example, we considered an RCT of a MI-based behavioral intervention. In this setting we argued Assumption 1.8 is more plausible than Assumption 1.7 and our proposed methods of estimating $ACE(cc)$ are more appropriate than the proposal of Schochet and Chiang [2011].

The proposed methodology is based on a novel set of identifying assumptions and consists of two corresponding estimators of $ACE(cc)$: $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$. In Section 1.3 we provide conditions on the observed data that must hold under the identifying assumptions, hence providing a means to validate the assumptions. We also described two extensions of the proposal of Schochet and Chiang [2011]: (1) we extend their proposed estimator to the setting where providers and subjects in the untreated arm of the study cannot crossover to treatment, and (2) we take a ML approach instead of a MOM approach to estimate $ACE(cc)$ under Assumptions 1.1 - 1.6 and 1.7.

Our simulation studies showed that the estimator for $ACE(cc)$ proposed by Schochet & Chiang, $ACE(cc)^{\text{SC}}_{\text{MOM}}$, has desired asymptotic behavior when its underlying assumptions hold (Assumptions 1.1 - 1.6 and 1.7); similarly, $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ have desired asymptotic behavior under Assumptions 1.1 - 1.6 and 1.8. However, the desirable asymptotic behavior of $ACE(cc)^{\text{SC}}_{\text{MOM}}$, $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ was not robust to violations of the estimators' underlying assumptions. Specifically, when Assumptions 1.1 - 1.6 and 1.8 hold but Assumption 1.7 does not hold, our simulations showed that $ACE(cc)^{\text{SC}}_{\text{MOM}}$ can be badly biased and its estimated 95% confidence interval can have poor coverage. Similarly, when instead Assumption 1.7 holds but Assumption 1.8 does not hold, $ACE(cc)_{\text{MOM}}$ and $ACE(cc)_{\text{ML}}$ can be biased and 95% confidence intervals can have poor coverage. The simulation study suggests that the scientific researcher should carefully consider which assumptions hold when choosing a method of estimating $ACE(cc)$.

In an example RCT comparing subjects assigned to MI or no counseling, we found that Assumptions 1.1 - 1.6 and 1.8 appeared to be valid. Although, we did not find statistical evidence to support a causal effect of MI. In the example data, there were a total of 17 therapists and covariate information on each therapist and patient was available. In the next two chapters, we extend the methods proposed here to consider covariates in modeling provider-subject compliance types and subject outcomes as well as clustering effects in modeling compliance types and outcomes within providers.

Chapter 2

ESTIMATING CONDITIONAL CAUSAL EFFECTS OF TREATMENT IN RANDOMIZED CONTROLLED TRIALS

In Chapter 1 we proposed a causal estimand that represents the marginal causal effect of MI based on potential outcomes [Neyman, 1990, Rubin, 1974, 1978] and principal stratification [Frangakis and Rubin, 2002]. The effect of MI may differ depending on therapist or patient pre-treatment covariates, however covariates are ignored in the methods of Chapter 1. For example, the targeted behavior for the MI treatment is substance abuse in Krupski et al. [2012]. Drug use severity is known to be related to patient's willingness to change as well as their substance abuse outcomes. Specifically, one would expect fewer compliers and lesser effect of MI among subjects that engage in lower risk substance abuse compared to subjects that engage in higher risk abuse. High risk patients more readily accept that they have a problem, hence are more likely to comply with treatment, and face more serious consequences, hence treatment is more likely to impact outcomes.

Covariates allow one to estimate different average treatment effects for subpopulations. When covariates are good predictors of compliance status, incorporating the covariates into the statistical model allows for a more precise partitioning of the sample with respect to subpopulations of interest. Generally, for such covariates, assignment is highly correlated with treatment receipt conditional on the covariates, which allows for more precise estimation of treatment effects in subpopulations of interest [Imbens and Rubin, 1997].

In this chapter we extend the methodology of Chapter 1 by incorporating covariates to define and propose estimators of a conditional causal effect of treatment. For estimation, we consider two approaches: (1) model the observed data (treatment assignment, observed compliance behavior and outcome) as a function of covariates, or (2) model the complete

data (treatment assignment, unobserved compliance type and outcome) as a function of covariates.

2.1 Definition of the Conditional Causal Effect of Treatment

We use similar notation to Chapter 1. First, consider the observed data. Suppose there are n provider-subject pairs and let i index pairs. Denote treatment assignment for the ith pair by Z_i ($Z_i = 1$ if the pair is assigned to treatment and $Z_i = 0$ otherwise). For the ith pair, denote the provider's observed compliance by D_i^{P} and the subject's observed compliance by D_i^{S}; $D_i^{\mathrm{P}} = 1$ if the provider follows the treatment protocol and $D_i^{\mathrm{P}} = 0$ otherwise; $D_i^{\mathrm{S}} = 1$ if the subject takes the treatment and $D_i^{\mathrm{S}} = 0$ otherwise. Denote the subject's binary outcome in the ith pair by Y_i. Denote the covariates of the ith provider-subject pair by $X_i = \begin{pmatrix} X_i^S \\ X_i^P \end{pmatrix}$, where X_i^S is a column vector of covariates collected on the subject and X_i^P is a column vector of covariates collected on the provider.

Next, consider the unobserved data. Let $\mathcal{N}$ denote the set of all n-dimensional column vectors of zeros and ones indicating all possible treatment assignments; hence $|\mathcal{N}| = 2^n$ and $\mathbf{z} \in \mathcal{N}$ represents one possible treatment assignment to the n provider-subject pairs. For the ith provider-subject pair, define the potential compliance under $\mathbf{z}$ by $D_i^{\mathrm{P}}(\mathbf{z})$ and $D_i^{\mathrm{S}}(\mathbf{z})$ for $\mathbf{z} \in \mathcal{N}$, and define the potential subject outcome under assignment $\mathbf{z}$ by $Y_i(\mathbf{z}) \equiv Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z}))$ for $\mathbf{z} \in \mathcal{N}$ where $\mathbf{D}^{\mathrm{P}}(\mathbf{z}) = (D_1^{\mathrm{P}}(\mathbf{z}), ..., D_n^{\mathrm{P}}(\mathbf{z}))$ and $\mathbf{D}^{\mathrm{S}}(\mathbf{z}) = (D_1^{\mathrm{S}}(\mathbf{z}), ..., D_n^{\mathrm{S}}(\mathbf{z}))$. There are 16 subgroups of provider-subject pairs defined by contrasting potential compliance under $z_i = 1$ (treatment) and $z_i = 0$ (no treatment) with $\mathbf{z}_{-i}$ otherwise fixed. Denote the compliance type for the ith provider-subject pair by C_i, which takes on values in $\mathcal{C} = \{st : s, t \in \{c, a, n, d\}\}$.

In Chapter 1, the difference in expected potential outcomes for provider-subject pairs with compliance type *cc* defines the causal effect of treatment under assignment $\mathbf{z}$:

$$P(Y_i(Z_1 = z_1, ..., Z_i = 1, ..., Z_n = z_n) = 1 \mid C_i = cc) \\ - P(Y_i(Z_1 = z_1, ..., Z_i = 0, ..., Z_n = z_n) = 1 \mid C_i = cc).$$

However, a conditional causal effect of treatment may also be of interest among provider-subject pairs of type cc for covariate values x, which we define as follows:

$$P(Y_i(Z_1 = z_1, ..., Z_i = 1, ..., Z_n = z_n) = 1 \mid C_i = cc, X_i = x) \\ - Y_i(Z_1 = z_1, ..., Z_i = 0, ..., Z_n = z_n \mid C_i = cc, X_i = x),$$

In the following sections, we propose assumptions for defining a conditional causal estimand of interest as well as identify it.

2.2 A Conditional Causal Estimand of Interest: $ACE(cc|x)$

Three assumptions are required to define a causal estimand of interest: Randomization, the Stable Unit Treatment Value Assumption (SUTVA), and Independent and Identical Distribution (iid) of the provider-subject pairs.

Assumption 2.1 (Randomization) *For* $i = 1, ..., n$,

$$Z_i \perp\!\!\!\perp D_i^{\mathrm{P}}(\mathbf{z}), D_i^{\mathrm{S}}(\mathbf{z}), Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})), X_i \text{ for } \mathbf{z} \in \mathcal{N}.$$

Assumption 2.2 (SUTVA) *For all* $\mathbf{z}$ *and* $i = 1, ..., n$,

$$D_i^{\mathrm{P}}(\mathbf{z}) = D_i^{\mathrm{P}}(z_i), D_i^{\mathrm{S}}(\mathbf{z}) = D_i^{\mathrm{S}}(z_i) \text{ and } Y_i(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})) = Y_i(z_i, D_i^{\mathrm{P}}(z_i), D_i^{\mathrm{S}}(z_i)).$$

The observed data $(Y_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}})$ *are related to the potential values as follows:*

$$D_i^{\mathrm{P}} = \sum_{z \in \{0,1\}} D_i^{\mathrm{P}}(z) \mathbb{1}(Z_i = z), \; D_i^{\mathrm{S}} = \sum_{z \in \{0,1\}} D_i^{\mathrm{S}}(z) \mathbb{1}(Z_i = z), \quad \text{and}$$

$$Y_i = \sum_{z \in \{0,1\}} \sum_{u \in \{0,1\}} \sum_{v \in \{0,1\}} Y_i(z, u, v) \mathbb{1}(Z_i = z, D_i^{\mathrm{P}}(z) = u, D_i^{\mathrm{S}}(z) = v).$$

Assumption 2.3 (iid) *For* $i = 1, ..., n$ *and* $x \in support(X) \equiv \mathcal{S}$,

$$Y_i \mid Z_i = z, C_i = st, X_i = x \overset{iid}{\sim} Bernoulli\big(\eta_{stzx}\big) \text{ for } st \in \mathcal{C}, \; z \in \{0, 1\} \text{ and}$$

$$C_i \mid X_i = x \sim Multinomial_{16}\big(1, \pi = (\pi_{stx} : st \in \mathcal{C})\big).$$

Next, we assume provider-subject pairs are iid *conditional* on covariate values x, which relaxes the (unconditional) iid assumption of Chapter 1.[1] In Chapter 1, the iid assumption used to define the average causal treatment effect may be too strong. In the MI example, patient outcomes and compliance type both depend on drug use severity (which, among other things, may represent the type of substance that is abused, such as alcohol vs. heroin). However MI is expected to be more effective among subjects with higher drug use severity (e.g. heroin users) compared to those with lower severity (e.g. alcohol users).

Under Assumptions 2.1 - 2.3, define the conditional causal estimand of interest, *the average conditional causal effect of treatment among provider-subject pairs that comply with assignment conditional on* x, denoted by $ACE(cc|x)$, as follows.

Definition 2.1 ($ACE(cc|x)$)

$$\begin{aligned} ACE(cc|x) &= P(Y_i = 1 \mid Z_i = 1, C_i = cc, X_i = x) - P(Y_i = 1 \mid Z_i = 0, C_i = cc, X_i = x) \\ &= \eta_{cc1x} - \eta_{cc0x}. \end{aligned}$$

Now that we have defined the conditional causal estimand of interest, there are two important considerations. First, the parameters of interest, η_{cc1x} and η_{cc0x}, cannot be identified from the observed data distribution without further assumptions. Second, the conditional distributions defined in Assumption 2.3 have potentially infinite-dimensional parameter spaces. For example, if there is only one, continuous-valued covariate, $|\mathcal{S}|$ is infinity and hence so is $|(\eta_{stzx} : st \in \mathcal{C}, z \in \{0,1\})|$ and $|\pi|$. In order to estimate $ACE(cc|x)$, we will consider two modeling approaches in Section 2.3. We now consider four additional assumptions that identify $ACE(cc|x)$, which are analogous to those made in Chapter 1.

Assumption 2.4 (Monotonicity) *For* $i = 1, ..., n$,

$$D_i^{\mathrm{P}}(1) \geq D_i^{\mathrm{P}}(0) \quad \textit{and} \quad D_i^{\mathrm{S}}(1) \geq D_i^{\mathrm{S}}(0).$$

Hence no provider-subject pairs have compliance type $dc, da, dn, dd, cd, ad,$ *or* nd.

[1] We further relax the iid assumption of Chapter 1 to account for nesting of subjects within providers in Chapter 3.

Assumption 2.5 (Additional Compliance Type Restrictions) *For* $i = 1, ..., n$,

$$\textit{if } \; D_i^{\mathrm{P}}(1) = D_i^{\mathrm{P}}(0) \;\; \textit{then } \; D_i^{\mathrm{S}}(1) = D_i^{\mathrm{S}}(0).$$

Hence no provider-subject pairs have compliance type ac or nc.

Assumption 2.6 (Stochastic Exclusion Restrictions)

$$\eta_{st1x} = \eta_{st0x} \equiv \eta_{stx} \textit{ for } st \in \{aa, an, na, nn\} \textit{ and } x \in \mathcal{S}.$$

Assumption 2.7 (Additional Stochastic Exclusion Restrictions)

$$\eta_{cn0x} = \eta_{nnx} \equiv \eta_{\bullet nx} \;\; \textit{and} \;\; \eta_{ca1x} = \eta_{aax} \equiv \eta_{\bullet ax} \textit{ for } x \in \mathcal{S}.$$

The "$\bullet$" in $\eta_{\bullet nx}$ and $\eta_{\bullet ax}$ denotes the provider-specific contribution to the compliance type may vary (e.g. first letter of the compliance type is c or n) but the subject-specific contribution does not (e.g. second letter is n).

Proposition 2.1 *Under Assumptions 2.1 - 2.7,* $ACE(cc|x)$ *is identifiable.*

The proof of Proposition 2.1 follows from the proof of Proposition 1.2 in Appendix A, where each probability is replaced with a conditional probability on x.

2.3 Proposed Estimators for $ACE(cc|x)$

Under Assumptions 2.1 - 2.7, the $ACE(cc|x)$ is identified from the observed data. However, the number of parameters of the model may be infinite depending on the cardinality of the support of the covariates. In order to estimate $ACE(cc|x)$, we take two approaches: in Section 2.3.1 we assume a parametric model for the observed data, while in Section 2.3.2 we assume a parametric model for the complete data. Either approach can easily handle a finite number[2] of continuous and/or discrete covariates, we contrast the approaches in Section 2.3.3.

[2]In both approaches we develop estimators that rely on large sample asymptotic theory and hence presume the setting where where the number of covariates is much smaller than the number of observations ($p << n$).

2.3.1 Parametric Model for Observed Data

Our first approach is to assume a parametric model on the observed data $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ conditional on X_i with finite dimensional parameter θ. We model $P(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i|X_i)$ with a single Multinomial logistic regression, however one could also specify a multi-part model (i.e. decompose $P(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i|X_i)$ and model its components with separate models). For example, $P(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i|X_i) = P(Y_i|Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, X_i)P(D_i^{\mathrm{P}}, D_i^{\mathrm{S}}|Z_i, X_i)P(Z_i)$, so one could alternately specify a functional form for $P(Y_i|Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, X_i)$ and $P(D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, |Z_i, X_i)$, separately.

Define the parameters of the Multinomial logistic regression model by β_{zuvy} for $z, u, v, y \in \{0, 1\}$ with β_{0000} equal to the zero vector. The finite dimensional parameter of the model is $\theta = (\beta_{0000}, ..., \beta_{1111})$. For a fixed x, we model $P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid X_i = x) \equiv \rho_{zuvy}$ by:

$$\rho_{zuvy} = \frac{e^{\mathsf{X}\beta_{zuvy}}}{\sum_{z,u,v,y\in\{0,1\}} e^{\mathsf{X}\beta_{zuvy}}} \text{ for } z, u, v, y \in \{0, 1\} \tag{2.1}$$

where X denotes rows from a design matrix that correspond to x. In general, $\rho = (\rho_{0000}, ..., \rho_{1111})$ defines a mapping of (θ, x).

By Proposition 2.1, the following relationships between the complete data model parameters, η_{cc1x} and η_{cc0x}, and the observed data Multinomial distribution parameter $(\rho_{0000}, ..., \rho_{1111})$ hold:

$$\eta_{cc0x} = \frac{(\rho_{0001}\rho_{100\bullet} + \rho_{1001}\rho_{010\bullet})\rho_{1\bullet\bullet\bullet} - (\rho_{1001}\rho_{100\bullet} + \rho_{1001}\rho_{110\bullet})\rho_{0\bullet\bullet\bullet}}{(\rho_{000\bullet}\rho_{100\bullet} + \rho_{100\bullet}\rho_{010\bullet})\rho_{1\bullet\bullet\bullet} - (\rho_{100\bullet}\rho_{100\bullet} + \rho_{100\bullet}\rho_{110\bullet})\rho_{0\bullet\bullet\bullet}} \tag{2.2}$$

$$\eta_{cc1x} = \frac{(\rho_{1111}\rho_{011\bullet} + \rho_{0111}\rho_{101\bullet})\rho_{0\bullet\bullet\bullet} - (\rho_{0111}\rho_{011\bullet} + \rho_{0111}\rho_{001\bullet})\rho_{1\bullet\bullet\bullet}}{(\rho_{111\bullet}\rho_{011\bullet} + \rho_{011\bullet}\rho_{101\bullet})\rho_{0\bullet\bullet\bullet} - (\rho_{011\bullet}\rho_{011\bullet} + \rho_{011\bullet}\rho_{001\bullet})\rho_{1\bullet\bullet\bullet}}, \tag{2.3}$$

where $\rho_{zuv\bullet} = \sum_{y=0,1} \rho_{zuvy}$ and $\rho_{z\bullet\bullet\bullet} = \sum_{u=0,1} \sum_{v=0,1} \sum_{y=0,1} \rho_{zuvy}$.

To condense the expression for $ACE(cc)$ and its variance, we reduce the number of summations in the expressions for η_{cc0x} and η_{cc1x}, above, with the following reparametrization:

$$\begin{array}{llll} \mu_1 = \rho_{1111} & \mu_5 = \rho_{0\bullet\bullet\bullet} & \mu_9 = \rho_{100\bullet} & \mu_{13} = \rho_{0011} \\ \mu_2 = \rho_{011\bullet} & \mu_6 = \rho_{0\bullet1\bullet} & \mu_{10} = \rho_{1001} & \mu_{14} = \rho_{0101} \\ \mu_3 = \rho_{0111} & \mu_7 = \rho_{1\bullet\bullet\bullet} & \mu_{11} = \rho_{010\bullet} & \mu_{15} = \rho_{1101} \\ \mu_4 = \rho_{101\bullet} & \mu_8 = \rho_{0001} & \mu_{12} = \rho_{1\bullet0\bullet} & \mu_{16} = \rho_{1011} \end{array}.$$

In other words, $\mu = (\mu_1, ..., \mu_{16})$ defines a mapping of ρ. Hence $ACE(cc|x)$ is defined by the following composite function:

$$\begin{aligned} ACE(cc|x) &\equiv (f \circ \mu \circ \rho)(\theta, x) \\ &= \frac{\mu_1\mu_2\mu_5 + \mu_3\mu_4\mu_5 - \mu_3\mu_6\mu_7}{\mu_7\mu_2\mu_5 - \mu_{12}\mu_2\mu_5 - \mu_2\mu_6\mu_7} - \frac{\mu_8\mu_9\mu_7 + \mu_{10}\mu_{11}\mu_7 - \mu_{10}\mu_{12}\mu_5}{\mu_5\mu_9\mu_7 - \mu_6\mu_9\mu_7 - \mu_9\mu_{12}\mu_5}. \end{aligned} \quad (2.4)$$

In order to estimate $ACE(cc|x)$, we first estimate θ. Assuming the model is the Multinomial logistic regression defined in (2.1), standard software can be used to estimate θ (e.g. the `multinom` function of the `nnet` package in R).[3] Denote the estimates for θ by $\hat{\theta}$. Define the estimator of $ACE(cc|x)$ based on an observed data parametric model by $ACE(cc|x)_{\text{observed}} = (f \circ \mu \circ \rho)(\hat{\theta}, x)$.

A estimator of the variance of $ACE(cc|x)_{\text{observed}}$ can be obtained in order to construct confidence intervals for $ACE(cc|x)_{\text{observed}}$. First, obtain the variance-covariance matrix for $\hat{\theta}$, and denote it by $V(\hat{\theta})$. An estimator of $V(\hat{\theta})$ is the inverse of the negative Hessian of the log likelihood of the parametric model evaluated at $\hat{\theta}$. For the Multinomial logistic regression defined in (2.1), the estimate of $V(\hat{\theta})$ is easily obtained from the statistical software used to fit the Multinomial logistic regression (e.g. by applying the `vcov` function to the fitted regression model object in R).[4] By the delta method, an estimator for the variance of $ACE(cc|x)_{\text{observed}}$

[3]Note that if we had instead specified a multi-part model, standard software could still be used to fit the models.

[4]However, the variance-covariance matrix is not as easily obtained for a multi-part model.

is:

$$\widehat{Var}(ACE(cc|x)_{\text{observed}}) = \Big(\frac{\partial(f\circ\mu\circ\rho)(\theta,x)}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big)\; V(\hat{\theta})\; \Big(\frac{\partial(f\circ\mu\circ\rho)(\theta,x)}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big)^{\mathsf{T}}$$
$$= \Big(\frac{\partial f}{\partial\mu}\Big|_{\mu=\hat{\mu}}\Big)\Big(\frac{\partial\mu}{\partial\rho}\Big|_{\rho=\hat{\rho}}\Big)\Big(\frac{\partial\rho}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big)\; V(\hat{\theta})\; \Big(\frac{\partial\rho}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big)^{\mathsf{T}}\Big(\frac{\partial\mu}{\partial\rho}\Big|_{\rho=\hat{\rho}}\Big)^{\mathsf{T}}\Big(\frac{\partial f}{\partial\mu}\Big|_{\mu=\hat{\mu}}\Big)^{\mathsf{T}},$$

where $\hat{\rho} = \rho(\hat{\theta}, x)$ and $\hat{\mu} = \mu(\hat{\rho})$. From equation (2.4), it is evident that $\frac{\partial f}{\partial\mu}|_{\mu=\hat{\mu}}$ is lengthy but simple to compute, and so we do not reproduce it here, however it is easily obtained via a computer algebra system (e.g. in R we utilized YACAS via the `Ryacas` package). $\frac{\partial\mu}{\partial\rho}|_{\rho=\hat{\rho}}$ is a 16×16 matrix of zeros and ones whose elements are easily intuited. The rows and columns of correspond to the elements of the vectors μ and ρ, respectively. For example, the matrix element corresponding to μ_1 and ρ_{1111} is 1 but all other elements of the row corresponding to μ_1 are zero). Finally, $\frac{\partial\rho}{\partial\theta}\big|_{\theta=\hat{\theta}}$ is given by the following:

$$\frac{\partial\rho}{\partial\theta}\Big|_{\theta=\hat{\theta}} = -\frac{1}{\big(\sum_{z,u,v,y} e^{\mathsf{X}\hat{\theta}_{zuvy}}\big)^2}\begin{pmatrix}1\\ e^{\mathsf{X}\hat{\theta}_{0001}}\\ \vdots\\ e^{\mathsf{X}\hat{\theta}_{1111}}\end{pmatrix}\begin{pmatrix}\mathsf{X}e^{\mathsf{X}\hat{\theta}_{0001}} & \cdots & \mathsf{X}e^{\mathsf{X}\hat{\theta}_{1111}}\end{pmatrix}$$
$$+\frac{1}{\sum_{z,u,v,y} e^{\mathsf{X}\hat{\theta}_{zuvy}}}\begin{pmatrix}\mathbf{0}^{\mathsf{T}} & \cdots & \mathbf{0}^{\mathsf{T}}\\ \hline \mathsf{X}e^{\mathsf{X}\hat{\theta}_{0001}} & & 0\\ & \ddots & \\ 0 & & \mathsf{X}e^{\mathsf{X}\hat{\theta}_{1111}}\end{pmatrix}.$$

2.3.2 Parametric Model for the Complete Data

A second approach is to assume a parametric model on the complete data $(Z_i, D^{\mathrm{P}}, D^{\mathrm{S}}, C_i, Y_i)$ conditional on X_i with finite dimensional parameter θ. Since treatment assignment (Z_i) is

randomized,

$$\begin{aligned} P(Y_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, C_i, Z_i \mid X_i) &= P(Y_i \mid D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, C_i, Z_i, X_i) P(D_i^{\mathrm{P}}, D_i^{\mathrm{S}} \mid C_i, Z_i, X_i) P(C_i \mid Z_i, X_i) P(Z_i \mid X_i) \\ &= P(Y_i \mid C_i, Z_i, X_i) P(D_i^{\mathrm{P}}, D_i^{\mathrm{S}} \mid C_i, Z_i, X_i) P(C_i \mid X_i) P(Z_i \mid X_i). \end{aligned}$$

The observed compliance $(D_i^{\mathrm{P}}, D_i^{\mathrm{S}})$ is determined by compliance type (C_i) and treatment assignment (Z_i), hence $P(D_i^{\mathrm{P}}, D_i^{\mathrm{S}} \mid C_i, Z_i, X_i)$ is either zero or one. Since treatment is binary and randomized, $P(Z_i \mid X_i) = P(Z_i) \equiv \xi$ for all i. This leaves two parts to model: $P(Y_i = 1 \mid C_i = st, Z_i = z, X_i = x)$ and $P(C_i = st \mid X_i = x)$.

Define the parameter of the complete data distribution model by $\theta = (\beta, \zeta, \xi)$ where $\beta = (\beta_{cc0}, \beta_{cc1}, \beta_{cn1}, \beta_{ca0}, \beta_{an}, \beta_{na}, \beta_{\bullet n}, \beta_{\bullet a})$ and $\zeta = (\zeta_{cc}, \zeta_{cn}, \zeta_{ca}, \zeta_{an}, \zeta_{na}, \zeta_{nn}, \zeta_{aa})$ with ζ_{cc} equal to the zero vector. For a fixed x, we model $P(Y_i = 1 \mid C_i = st, Z_i = z, X_i = x)$ with a conditional logistic regression model:

$$\eta(x; \beta_\star) = \frac{e^{\mathsf{X}_\eta \beta_\star}}{1 + e^{\mathsf{X}_\eta \beta_\star}} \text{ for } \star \in \{cc0, cc1, cn1, ca0, an, na, \bullet n, \bullet a\} \tag{2.5}$$

where X_η denotes a row from a design matrix that corresponds to x. We model $P(C_i = st \mid X_i = x)$ with a conditional Multinomial logistic regression model:

$$\pi(x; \zeta_{st}) = \frac{e^{\mathsf{X}_\pi \zeta_{st}}}{\sum_{st \in \mathcal{C}_7} e^{\mathsf{X}_\pi \zeta_{st}}} \text{ for } st \in \mathcal{C}_7 \equiv \{cc, cn, ca, an, na, nn, aa\} \tag{2.6}$$

where X_π denotes a row from a design matrix that corresponds to x (which may differ from X_η since the multi-part model construction allows the function form of each part to be specified separately). The complete data likelihood is

$$L = \prod_{i=1}^{n} f(C_i, Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i \mid X_i, \theta)$$

where $f(\cdot)$ is the density function of a multinomial distribution where counts depend on $(C_i, Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ and cell probabilities depend on (X_i, θ).

We use the EM algorithm to estimate $ACE(cc|x)$ and its variance. In the kth E-step, we compute $E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$, the conditional expectation of the complete data log-likelihood given the observed data under the current parameter value. In the kth M-step,

the update $\theta^{(k+1)}$ maximizes the computed expectation. To obtain the maximizer, usually the derivative $\frac{\partial}{\partial\theta}E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ is set equal to zero, and the update $\theta^{(k+1)}$ is set to the maximizer. However in this setting no closed form solution exists. The ML solution could be found numerically, but it would be computationally intensive as it would require an iterative algorithm within each M-step. Hence we implement the generalized EM algorithm, or GEM-algorithm [Dempster et al., 1977]. In the GEM algorithm, the update $\theta^{(k+1)}$ need only satisfy

$$E_{\theta^{(k+1)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \geq E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \tag{2.7}$$

as opposed to maximizing $E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ [Dempster et al., 1977]. Specifically, Lange [1995] showed that the solution of the first step of the Newton-Raphson procedure for obtaining a root of $E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) = 0$ satisfies (2.7). Using the update based on one iteration of Newton Raphson is known as the EM gradient algorithm, a special case of the GEM-algorithm. Details of implementing the EM gradient algorithm are given in Appendix B. The E- and M-steps are repeated until convergence. Multiple starting values for $\theta^{(0)}$ are tried, the final update that corresponds to the highest converged likelihood value is considered the ML estimate that corresponds to the global maximum, which we denote by $\hat{\theta}$. The ML estimate for $ACE(cc|x)$, denoted by $ACE(cc|x)_{\text{complete}}$ is thus

$$ACE(cc|x)_{\text{complete}} = \eta(x; \hat{\beta}_{cc1}) - \eta(x; \hat{\beta}_{cc0}).$$

An asymptotically correct estimate of the variance-covariance matrix of $\hat{\theta}$ can be estimated from the inverse of the observed data information matrix, $I_{\mathbf{n}}$. Since computing the gradient or Hessian of the observed data likelihood is cumbersome, we use Louis's formula [Louis, 1982] to obtain $I_{\mathbf{n}}$:

$$I_{\mathbf{n}} \;=\; E_{\hat{\theta}}\left(- \left.\frac{\partial^2(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial\theta\partial\theta^{\intercal}}\right|_{\theta=\hat{\theta}}\right) - Cov_{\hat{\theta}}\left(\left.\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial\theta}\right|_{\theta=\hat{\theta}}\right).$$

(The explicit form of $I_{\mathbf{n}}$ is given in Appendix B.) By the delta method, an asymptotically consistent estimate of the variance of $ACE(cc|x)_{\text{complete}}$ is

$$\widehat{Var}(ACE(cc|x)_{\text{complete}})$$

$$= \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\intercal} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \frac{\mathsf{X}_\eta e^{\mathsf{X}_\eta \hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}_\eta \hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \frac{\mathsf{X}_\eta e^{\mathsf{X}_\eta \hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}_\eta \hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^{\intercal} I_{\mathbf{n}}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \frac{\mathsf{X}_\eta e^{\mathsf{X}_\eta \hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}_\eta \hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \frac{\mathsf{X}_\eta e^{\mathsf{X}_\eta \hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}_\eta \hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Further details regarding the derivation of $\widehat{Var}(ACE(cc|x)_{\text{complete}})$ are given in Appendix B.

2.3.3 Contrasting Parametric Models on Observed vs. Complete Data

In Sections 2.3.1 and 2.3.2 we proposed two approaches to estimating $ACE(cc|x)$. We now consider differences between the approaches as well as the relationship between the observed and complete data generating models. Since we used θ to define parameters in both Sections 2.3.1 and 2.3.2, to avoid confusion, let θ_{observed} denote the parameter in the observed data model and θ_{complete} denote the parameter in the complete data model.

Differences Between the Proposed Approaches

By assuming a parametric model for the observed data, we can use standard model checking procedures to assess model fit, such as a goodness of fit test. But, it is difficult to understand the relationship between covariates and the causal effect of treatment in this model. From (2.4) we can see that the relationship between θ_{observed} and $ACE(cc|x)$ is complex. It is not readily apparent how the covariates affect outcomes among provider-subject pairs of type *cc*. Another drawback of the approach is that the estimated conditional probabilities η_{cc0x} in (2.2) and η_{cc1x} in (2.3) may take on values outside of the interval $[0, 1]$. Moreover, the

parameters of the model are not guaranteed to be compatible with Assumptions 2.1 - 2.7. Under these assumptions, the inequalities derived in Section 1.3 of Chapter 1 must hold, where the probabilities are taken conditionally on x. However, there may be values of x for which these inequalities do not hold.

On the other hand, there is a clear relationship between θ_{complete} and $ACE(cc|x)$. β_{cc1} is interpreted as the differences in log odds of a successful outcome between subjects that have a 1-unit difference in X_i among treated subjects where the provider-subject pair is of type *cc*. β_{cc0} has a similar interpretation among untreated subjects. Hence the relative signs and magnitudes of β_{cc1} and β_{cc0} provide insight into the relationship between covariates and $ACE(cc|x)$. For example, suppose $\beta_{cc1} > 0$, $\beta_{cc0} < 0$ and X_i is a binary variable. This result would suggest that the treatment has a greater positive effect for subjects where $X_i = 1$ compared to subjects where $X_i = 0$. Furthermore, the parameters of the model are compatible with Assumptions 2.1 - 2.7. In the following section we discuss a relationship between the observed and complete data distributions that allows one to perform a goodness of fit test to assess model fit. Overall, the approach of 2.3.2 is preferable.

Observed Data Distribution from the Complete Data Parametric Model

The complete data for the ith provider-subject pair consists of the four binary variables, a discrete variable (compliance type) and a vector of covariates: $Z_i, D_i^{\text{P}}, D_i^{\text{S}}, Y_i, C_i$ and X_i, respectively. The conditional distribution of Z_i, Y_i, C_i given $X_i = x$ has a Multinomial distribution, $Multinomial_{28}(1, \rho^*)$, where the elements of ρ^* are indexed by $st \in \mathcal{C}_7$ and $z, y \in \{0, 1\}$. Note that Z_i and C_i completely determine D_i^{P} and D_i^{S}, and so the conditional distribution of $Z_i, D_i^{\text{P}}, D_i^{\text{S}}, Y_i, C_i$ given $X_i = x$ has a Multinomial distribution with 28×4 cells, however only 28 cells have nonzero probability.

Recall $\eta_{stzx} = P(Y_i = 1 \mid Z_i = z, C_i = st, X_i = x)$, $\pi_{stx} = P(C_i = st|X_i = x)$ and denote $P(Z_i|X_i) = \xi$ for all i. The elements of ρ^* are defined in Table 2.1a. The observed data for the ith provider-subject pair consists of the four binary variables and a vector of covariates: $Z_i, D_i^{\text{P}}, D_i^{\text{S}}, Y_i, X_i$. The conditional distribution of the observed data $Z_i, D_i^{\text{P}}, D_i^{\text{S}}, Y_i$ given

$X_i = x$ is $Multinomial_{16}(1, \rho)$. The cells of the observed data Multinomial distribution are combinations of the cells of the complete data Multinomial distribution, shown in Table 2.1b.

Assuming the parametric model defined by (2.5) and (2.6), $\eta_{stzx} = \eta(x; \beta_{stz})$ for $stz \in \{cc0, cc1, cn1, ca0\}$, $\eta_{nn0x} = \eta_{nn1x} = \eta_{cn0x} = \eta(x; \beta_{\bullet n})$, $\eta_{aa0x} = \eta_{aa1x} = \eta_{ca1x} = \eta(x; \beta_{\bullet a})$, $\eta_{an0x} = \eta_{an1x} = \eta(x; \beta_{an})$, $\eta_{na0x} = \eta_{na1x} = \eta(x; \beta_{na})$, and $\pi_{stx} = \pi(x; \zeta_{st})$ for $st \in \mathcal{C}_7$. Based on the connection between the conditional complete and observed distributions shown in Table 2.1, estimates of $P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid X_i = x)$ for $z, u, v, y \in \{0, 1\}$ and fixed x can be obtained from the parametric model defined by (2.5) and (2.6). In order to check the fit of the complete data parametric model, one could perform a chi-square goodness of fit test on the observed proportions of $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ and expected proportions given by the model and Table 2.1.

Complete Data Distribution from the Observed Data Parametric Model

By Proposition 2.1, the probabilities π_{stx} for $st \in \mathcal{C}_7$ and η_{stzx} for $st \in \mathcal{C}_7$ and $z \in \{0, 1\}$ can be expressed in terms of probabilities of the observed data $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ conditional on X_i. Hence, we can estimate $P(Z_i = z, C_i = st, Y_i = y \mid X_i = x)$ for $st \in \mathcal{C}_7$ and $z, y \in \{0, 1\}$ from the observed data. For example,

$$\pi_{nax} = P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1 \mid Z_i = 1, X_i = x) = \frac{P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1, Z_i = 1 \mid X_i = x)}{P(Z_i = 1 \mid X_i = x)}.$$

Since treatment is randomized, $P(Z_i = 1 \mid X_i = x) = P(Z_i = 1)$ can be estimated using proportion of provider-subject pairs assigned to $z = 1$. Assuming the model in (2.1), we can obtain an estimate of $P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1, Z_i = 1 \mid X_i = x)$ using the estimated coefficients.

From Table 2.1a, estimates for $P(Z_i = z, C_i = st, Y_i = y \mid X_i = x)$ for $st \in \mathcal{C}_7$ and $z \in \{0, 1\}$ can be obtained.

2.4 Simulation Studies

We perform two simulation studies where data is simulated to mimic findings in the MI literature. In both studies, Assumptions 2.1 - 2.7 hold. In the first study, a parametric

Table 2.1: Definition of the probability vectors of the (conditional) Multinomial distributions that gives rise to the (a) complete data and (b) observed data.

st	ρ^*_{st00}	ρ^*_{st01}	ρ^*_{st10}	ρ^*_{st11}
cc	$(1-\xi)\pi_{ccx}(1-\eta_{cc0x})$	$(1-\xi)\pi_{ccx}\eta_{cc0x}$	$\xi\pi_{ccx}(1-\eta_{cc1x})$	$\xi\pi_{ccx}\eta_{cc1x}$
cn	$(1-\xi)\pi_{cnx}(1-\eta_{\bullet nx})$	$(1-\xi)\pi_{cnx}\eta_{\bullet nx}$	$\xi\pi_{cnx}(1-\eta_{cn1x})$	$\xi\pi_{cnx}\eta_{cn1x}$
ca	$(1-\xi)\pi_{cax}(1-\eta_{ca0x})$	$(1-\xi)\pi_{cnx}\eta_{ca0x}$	$\xi\pi_{cax}(1-\eta_{\bullet ax})$	$\xi\pi_{cax}\eta_{\bullet ax}$
an	$(1-\xi)\pi_{anx}(1-\eta_{anx})$	$(1-\xi)\pi_{anx}\eta_{anx}$	$\xi\pi_{anx}(1-\eta_{anx})$	$\xi\pi_{anx}\eta_{anx}$
na	$(1-\xi)\pi_{nax})(1-\eta_{nax})$	$(1-\xi)\pi_{nax}\eta_{nax}$	$\xi\pi_{nax}(1-\eta_{nax})$	$\xi\pi_{nax}\eta_{nax}$
nn	$(1-\xi)\pi_{nnx}(1-\eta_{\bullet n}x)$	$(1-\xi)\pi_{nnx}\eta_{\bullet nx}$	$\xi\pi_{nnx}(1-\eta_{\bullet n})$	$\xi\pi_{nnx}\eta_{\bullet nx}$
aa	$(1-\xi)\pi_{aax}(1-\eta_{\bullet ax})$	$(1-\xi)\pi_{aax}\eta_{\bullet ax}$	$\xi\pi_{aax}(1-\eta_{\bullet ax})$	$\xi\pi_{aax}\eta_{\bullet ax}$

(a) Definition of the probability vector ρ^* of the conditional 28-celled Multinomial distribution that gives rise to the complete data Z_i, C_i, Y_i given $X_i = x$.

z	y	ρ_{z00y}	ρ_{z01y}	ρ_{z10y}	ρ_{z11y}
0	0	$\sum_{st\in\{cc,cn,nn\}} \rho^*_{st00}$	$\sum_{st\in\{ca,na\}} \rho^*_{st00}$	ρ^*_{an10}	ρ^*_{aa00}
0	1	$\sum_{st\in\{cc,cn,nn\}} \rho^*_{st01}$	$\sum_{st\in\{ca,na\}} \rho^*_{st01}$	ρ^*_{an11}	ρ^*_{aa01}
1	0	ρ^*_{nn10}	ρ^*_{na10}	$\sum_{st\in\{cn,an\}} \rho^*_{st10}$	$\sum_{st\in\{cc,ca,aa\}} \rho^*_{st10}$
1	1	ρ^*_{nn11}	ρ^*_{na11}	$\sum_{st\in\{cn,an\}} \rho^*_{st11}$	$\sum_{st\in\{cc,ca,aa\}} \rho^*_{st11}$

(b) Definition of the probability vector ρ of the conditional 16-celled Multinomial distribution that gives rise to the observed data $Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i$ given $X_i = x$.

model defined by (2.5) and (2.6) holds. In the second study a parametric model defined by (2.1) holds.

Two covariates are simulated: (1) subject's drug use severity and (2) an indicator for whether the therapist is a non-routine primary care provider. Patient outcomes are known to be affected by drug use severity. For example in Krupski et al. [2012], the authors used the DAST-10 to measure drug use severity, a discrete scale from 0-10 where 0 indicates no health risk from drug abuse and 1-10 represent increasing levels of risk. One of the study inclusion criteria for enrollment was a DAST-10 score of at least 1, and patients were block randomized to MI or usual care based on a DAST-10 score $\geq$3. Using the same study data, Dunn et al. [2015] found non-routine primary care providers compared to routine primary care providers were more likely to have higher MI fidelity scores. In the following simulation studies, let X_i^S be a binary indicator of whether the patient's baseline DAST-10 $\geq$ 3, and let X_i^P be an indicator of whether the therapist was a non-routine primary care provider.

2.4.1 Complete Data Parametric Model Simulations

First, we model $P(Y_i = 1 \mid C_i = st, Z_i = z, X_i = x)$ with the logistic regression models:

$$\eta\left(x = (x^S, x^P); \beta_\star = \begin{pmatrix} \beta_0 \\ \beta_S \end{pmatrix}\right) = \frac{e^{\beta_0 + \beta_S x^S}}{1 + e^{\beta_0 + \beta_S x^S}} \tag{2.8}$$

$$\text{for } \star \in \{cc0, cc1, cn1, ca0, an, na, \bullet n, \bullet a\}$$

and model $P(C_i = st \mid X_i = x)$ with the Multinomial logistic regression model:

$$\pi\left(x = (x^S, x^P); \zeta_{st} = \begin{pmatrix} \zeta_0 \\ \zeta_S \\ \zeta_P \end{pmatrix}\right) = \frac{e^{\zeta_0 + \zeta_S x^S + \zeta_P x^P}}{\sum_{st \in \mathcal{C}_7} e^{(1, x^S, x^P)\zeta_{st}}} \text{ for } st \in \mathcal{C}_7 \text{ with } \zeta_{cc} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.9}$$

The model parameter θ_{complete} was set to the following values. We fixed $\beta_{cc0} = \beta_{an} = \beta_{\bullet n} = \beta_{cn1} = (\log(0.5), 0)^\intercal$, $\beta_{cc1} = \beta_{na} = \beta_{\bullet a} = \beta_{ca0} = (\log(0.5), \log(3))^\intercal$ and the values of ζ according to Table 2.2. The values of β and ζ were chosen to simulate plausible relationships.

		π_{stx}			
st	ζ_{st}	$x=(0,0)$	$x=(1,0)$	$x=(0,1)$	$x=(1,1)$
cc	$(0,0,0)^\intercal$	0.203	0.282	0.227	0.312
nn	$(1,-1,-1)^\intercal$	0.552	0.282	0.227	0.115
cn	$(-1,-1,1)^\intercal$	0.075	0.038	0.227	0.115
an	$(-1,-1,1)^\intercal$	0.075	0.038	0.227	0.115
aa	$(-3,1,1)^\intercal$	0.010	0.038	0.031	0.115
ca	$(-3,1,1)^\intercal$	0.010	0.038	0.031	0.115
na	$(-1,1,-1)^\intercal$	0.075	0.282	0.031	0.115

Table 2.2: Value of ζ and values of $\pi_{stx} = P(C_i = st|X_i = x)$ in the simulation study of Section 2.4.

Table 2.3 shows the values of η_{ccz} for $z \in \{0,1\}$. The $ACE(cc|x)$ is only nonzero for patients with high severity drug use. Table 2.2 also shows the probability of each compliance type given x, π_{stx} for $st \in \mathcal{C}_7$. The comparative probabilities correspond to expected relationships. For example, the probabilities of compliance types *cc*, *ca* and *cn* are greater for therapists that were non-routine primary care providers compared to therapists that were not (e.g. comparing column 5 to column 3 in Table 2.3, the probabilities of type *cc*, *ca* and *cn* are all greater in column 5). In other words, the non-rountine primary care providers were more likely to comply with treatment assignment compared to the routine primary care providers.

	η_{cczx}			
z	$x=(0,0)$	$x=(1,0)$	$x=(0,1)$	$x=(1,1)$
1	0.333	0.600	0.333	0.600
0	0.333	0.333	0.333	0.333

Table 2.3: Value of η in the simulation study of Section 2.4.

2.4.2 Observed Data Parametric Model Simulations

Next, we model $P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid X_i = x)$ with the Multinomial logistic regression model:

$$\rho\left(x=(x^S,x^P);\beta_{zuvy}=\begin{pmatrix}\beta_0\\ \beta_S\\ \beta_P\end{pmatrix}\right)=\frac{e^{\beta_0+\beta_S x^S+\beta_P x^P}}{\sum_{z,u,v,y\in\{0,1\}} e^{(1,x^S,x^P)\beta_{zuvy}}} \text{ for } z,u,v,y\in\{0,1\}. \quad (2.10)$$

The observed data model parameter θ_{observed} is given in Table 2.4. The values of θ_{observed} were chosen so that the probabilities $P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid X_i = x)$ for $z, u, v, y \in \{0, 1\}$ are equal to those in Section 2.4.1.

2.4.3 Simulation Study Results

We simulated data from the model defined by (2.8) and (2.9) and the model defined by (2.10) for sample sizes of 100 to 100,000. Table 2.5 summarizes the results over 1000 Monte Carlo samples using the method proposed in Section 2.3.2 on data simulated from the model defined by (2.8) and (2.9). Table 2.6 summarizes the results over 1000 Monte Carlo samples using the method proposed in Section 2.3.1 on data simulated from the model defined by (2.10). From the tables, we can see that Bias and MSE decrease with sample size, and coverages of the 95% confidence interval approaches 95%.

β_{0000} = (0, 0, 0)	β_{1000} = (-0.408, -0.351, -0.691)
β_{0001} = (-0.693, 0.000, 0.000)	β_{1001} = (-1.101, -0.351, -0.691)
β_{0010} = (-2.281, 1.138, -0.125)	β_{1010} = (-2.408, 1.138, -0.691)
β_{0011} = (-2.974, 2.237, -0.125)	β_{1011} = (-3.101, 2.237, -0.691)
β_{0100} = (-2.408, -0.351, 1.309)	β_{1100} = (-1.714, -0.351, 1.309)
β_{0101} = (-3.101, -0.351, 1.309)	β_{1101} = (-2.408, -0.351, 1.309)
β_{0110} = (-4.408, 1.138, 1.309)	β_{1110} = (-1.313, 0.283, 0.454)
β_{0111} = (-5.101, 2.237, 1.309)	β_{1111} = (-2.006, 1.381, 0.454)

Table 2.4: Value of θ_{observed} in the simulation study of Section 2.4.

Results from implementing the method proposed in Section 2.3.2 assuming (2.8) and (2.9) on data simulated from the model defined by (2.10) were similar to those shown in Table 2.5. Likewise, results from implementing the method proposed in Section 2.3.1 assuming (2.10) on data simulated from the model defined by (2.8) and (2.9) were similar to those shown in Table 2.6. This suggests that in this simulation study, $ACE(cc|x)_{\text{observed}}$ and $ACE(cc|x)_{\text{complete}}$ are valid estimators regardless of whether the data are generated by the complete data parametric model defined by (2.5) and (2.6) or the observed parametric model defined by (2.1) as long as the specified model well approximates $P(Z_i = z, D_i^{\text{P}} = u, D_i^{\text{S}} = v, Y_i = y \mid X_i = x)$ for $z, u, v, y \in \{0, 1\}$. In fact, under the parameter values in Tables 2.2, 2.3 and 2.4 were chosen so that the simulated data have similar sample proportions of $(Z_i = z, D_i^{\text{P}} = u, D_i^{\text{S}} = v, Y_i = y)$ for all $z, u, v, y \in \{0, 1\}$, and so it is not surprising that we find similar results regardless of how the data were simulated.

2.5 Application to a RCT of MI

As a specific example, we applied our estimators of $ACE(cc|x)$ to data from an RCT of MI, where treatment access was not possible when $z = 0$ [Krupski et al., 2012]. A summary

	$x = (1,1)$			$x = (1,0)$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
500	0.11	0.081	0.857	-0.15	0.17	0.857
2000	-0.0014	0.017	0.940	0.0032	0.019	0.957
20000	-0.00018	0.0014	0.953	-0.00071	0.0014	0.947
	$x = (0,1)$			$x = (0,0)$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
500	-0.020	0.099	1.00	-0.32	0.21	1.00
2000	-0.021	0.027	0.954	-0.015	0.026	0.940
20000	0.00092	0.0023	0.956	0.00052	0.0022	0.950

Table 2.5: Estimates of $ACE(cc|x)$ and its variance for $x \in \{0,1\} \times \{0,1\}$ were obtained via the approach of Section 2.3.2 on data simulated from the model defined by (2.8) and (2.9).

of this study is given in Section 1.6 of Chapter 1. For the ith provider-subject pair, let Z_i denote treatment assignment, D_i^{P} denote a binary provider compliance indicator, D_i^{S} denote a binary subject compliance indicator, Y_i denote a binary indicator of whether the subject's drug use decreased at 3-months compared to baseline and X_i denote a binary indicator of the subject's baseline drug use severity ($X_i = 1$ for high severity, $X_i = 0$ for low severity).

The primary study aim is to estimate the causal effect of MI on patient outcomes. If a provider is unable to meet minimal proficiency standards regardless of training (i.e. comply with the treatment protocol), or a patient is unwilling to consider changing the their drug use no matter how skilled the provider (i.e. 'take' the treatment), it is unlikely that MI will have any impact on the patient outcome and an ITT analysis of the data will be biased toward a null intervention. Furthermore, the causal effect of MI is expected to differ by

	$x = (1,1)$			$x = (1,0)$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
500	-0.018	0.087	0.993	-0.0090	0.088	0.989
1000	0.0014	0.039	0.961	-0.0034	0.036	0.976
10000	-0.00036	0.0031	0.952	0.00076	0.0029	0.946
	$x = (0,1)$			$x = (0,0)$		
n	Bias	MSE	95% CI Coverage	Bias	MSE	95% CI Coverage
500	-0.0089	0.14	0.982	0.0043	0.30	0.979
1000	0.011	0.051	0.959	0.0026	0.089	0.970
10000	-0.0030	0.0045	0.950	-0.00060	0.0069	0.952

Table 2.6: Estimates of $ACE(cc|x)$ and its variance for $x \in \{0,1\} \times \{0,1\}$ were obtained via the approach of Section 2.3.1 on data simulated from the model defined by (2.10).

subgroups defined by a pre-treatment covariate, drug use severity.[5] To address the two-level noncompliance in this example, we assume Assumptions 2.1 - 2.7 from Section 2.2. We implement both the approaches proposed in Sections 2.3.1 and 2.3.2. Since treatment access was not possible when $z = 0$, we set $D_i^{\mathrm{P}} = 0$ and $D_i^{\mathrm{S}} = 0$ when $Z_i = 0$.[6]

[5]The causal effect of MI is also expected to differ by subgroups defined by whether or not the therapist was a routine primary care provider. However, since there are no provider covariates collected for the $z = 0$ arm of the study, we do not include the covariate in the parametric models.

[6]Both $ACE(cc|x)_{\mathrm{observed}}$ and $ACE(cc|x)_{\mathrm{complete}}$ and estimators of their variance have been modified in this section to accommodate the fact that $D_i^{\mathrm{P}} = 0$ and $D_i^{\mathrm{S}} = 0$ when $Z_i = 0$.

2.5.1 Assume a Parametric Model on the Complete Data

In order to implement the approach in Section 2.3.2, we assume the following parametric model:

$$\eta\left(x;\beta_\star=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\right)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\text{ for }\star\in\{cc0,cc1,cn1,\bullet n\} \tag{2.11}$$

$$\pi\left(x;\zeta_{st}=\begin{pmatrix}\zeta_0\\ \zeta_1\end{pmatrix}\right)=\frac{e^{\zeta_0+\zeta_1 x}}{\sum_{st\in\mathcal{C}_7}e^{(1,x)\zeta_{st}}}\text{ for }st\in\{cc,cn,nn\}\text{ with }\zeta_{cc}=\begin{pmatrix}0\\ 0\end{pmatrix}. \tag{2.12}$$

Note that compliance types *an*, *na* and *ca* are not possible when treatment access was not possible when $z=0$, and hence these types are not represented in the model.

Using the method proposed in Section 2.3.2, we estimate $ACE(cc|x=1)=0.05$ with an estimated variance of 0.04 and 95% CI: [-0.32, 0.42]. We estimate $ACE(cc|x=0)=-0.21$ with an estimated variance of 0.23 and 95% CI: [-1, 0.74]. Both estimates are not statistically significant at the $\alpha=0.05$ level. Nevertheless, under Assumptions 2.1 - 2.7, the point estimate for $ACE(cc|x=1)$ can be interpreted as a small, but not clinically meaningful, positive effect of MI on decreased drug use among subjects with high drug use severity. The point estimate of $ACE(cc|x=0)$ suggests there is a clinically relevant detrimental effect of MI among subjects with low drug use severity, yet uncertainty around this estimate is large.

2.5.2 Assume a Parametric Model on the Observed Data

In order to implement the approach in Section 2.3.1, we now assume the following parametric model:

$$\rho\left(x;\beta_{zuvy}=\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\right)=\frac{e^{\beta_0+\beta_1 x}}{\sum_{zuvy\in\mathcal{T}}e^{(1,x)\beta_{zuvy}}}\text{ for }zuvy\in\mathcal{T} \tag{2.13}$$

where $\mathcal{T} = \{0000, 0001\} \cup \{1uvy : u, v, y \in \{0, 1\}\}$. $\mathcal{T}$ represents all possible values of $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ when treatment access is not possible when $z = 0$ (since $D_i^{\mathrm{P}} = 0$ and $D_i^{\mathrm{S}} = 0$ when $Z_i = 0$).

Under Assumptions 2.1 - 2.7, the first two inequalities in (1.2) must hold, conditional on x. We check that this is the case with the fitted model, for all possible values of x:

$$P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 0 \mid Z_i = 0, X_i = 0) + P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 1 \mid Z_i = 1, X_i = 0) = 0.82 \leq 1,$$

$$P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 0 \mid Z_i = 1, X_i = 0) + P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 1 \mid Z_i = 0, X_i = 0) = 0.41 \leq 1,$$

$$P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 0 \mid Z_i = 0, X_i = 1) + P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 1 \mid Z_i = 1, X_i = 1) = 0.63 \leq 1,$$

$$P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 0 \mid Z_i = 1, X_i = 1) + P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0, Y_i = 1 \mid Z_i = 0, X_i = 1) = 0.54 \leq 1.$$

Note that the equality (1.4), conditional on x, holds due to the fact that the model (2.13) does not include cells for $(Z_i = z, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1, Y_i = y)$ and hence does not have positive probability for these cells. As in Section 1.6 of Chapter 1, two two provider-subject pairs were observed to have $D_i^{\mathrm{P}} = 0$, $D_i^{\mathrm{S}} = 1$ and $Z_i = 1$ but barely exceeded the threshold for making $D_i^{\mathrm{S}} = 1$. We attribute this to rounding error rather than a violation of (1.4) and Assumption 2.5. For these two pairs we have recoded $D_i^{\mathrm{S}} = 0$.

Using the method proposed in Section 2.3.1, we estimate $ACE(cc|x = 1) = 0.06$ with an estimated variance of 0.04 and 95% CI: [-0.33, 0.44]. We estimate $ACE(cc|x = 0) = -0.24$ with an estimated variance of 0.04 and 95% CI: [-0.65, 0.16]. Both estimates are not statistically significant at the $\alpha = 0.05$ level, but similar to those found in Section 2.5.1.

2.5.3 Goodness of Fit

To examine how well the model defined by (2.11) and (2.12) and model defined by (2.13) fit the data, we performed chi-square goodness of fit tests. The observed and model-based expected counts are given in Table 2.7. The test statistic for the model defined by (2.11) and (2.12) is 25.4 (p-value=.0006). This suggests that the data are not consistent with the model complete data parametric model. The test statistic for the model defined by (2.11) is 13.7

(p-value=0.057). Based on the test statistics, it appears that the observed data parametric model is more consistent with the data compared to the complete data parametric model in this example.

	$P(Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y)$							
	$z =$ 0	0	1	1	1	1	1	1
	$u =$ 0	0	0	0	1	1	1	1
	$v =$ 0	0	0	0	0	0	1	1
Counts	$y =$ 0	1	0	1	0	1	0	1
Observed	241	148	6	4	19	11	8	7
Expected based on (2.13)	272.3	124.3	6.1	3.5	18.9	10.1	5.4	3.3
Expected based on (2.11) & (2.12)	278.5	108.8	9.0	3.2	25.2	9.0	6.6	3.8

Table 2.7: Observed and model-based expected counts of $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ in the MI example dataset considered in Section 2.5. Two expected counts are computed: the first is based on the observed data parametric model defined by (2.13), the second is based on the complete data parametric model defined by (2.11) and (2.12).

Note that since we had only one binary covariate, we only had to check that the assumptions appear valid for $x = 0$ and $x = 1$ via the inequalities derived in Section 1.3 of Chapter 1. However, with additional, possibly continuous-valued covariates, there is greater potential for observed or new covariate values to violate the inequalities, which would suggest the fitted parametric model on the observed data is inconsistent with Assumptions 2.1 - 2.7.

2.6 Final Remarks

In this chapter we considered conditional causal effects of treatment, motivated by research of MI for problem drug use. In the motivating example, the patient's drug use severity and the therapist's professional status are expected to predict compliance type and the effect of MI is expected to differ for patient's with low vs. high drug use severity.

We proposed two approaches to estimating the average causal effect of treatment among

provider-subject pairs that comply with assignment, or $ACE(cc|x)$. In the first approach we assume a parametric model on the observed data, however warn that this approach may lead to a fitted model that is inconsistent with Assumptions 2.1 - 2.7. In Chapter 2.3.1, we modeled the observed data with a single Multinomial logistic regression, however note that much of a multi-part model could be used instead. In the second approach we assume a parametric model on the complete data, which is defined in terms of the parameters in Assumptions 2.1 - 2.7. Our simulation studies showed the proposed estimators $ACE(cc|x)_{\text{observed}}$ and$ACE(cc|x)_{\text{complete}}$ are well behaved asymptotically when the model is correctly specified, and under particular circumstances, they can also be valid when the model is not correctly specified. Specifically, when the probabilities $P(Z_i = z, D_i^{\text{P}} = u, D_i^{\text{S}} = v, Y_i = y)$ for $z, u, v, y \in \{0, 1\}$ are well approximated by the model.

In Section 2.3.3 we compared $ACE(cc|x)_{\text{observed}}$ and $ACE(cc|x)_{\text{complete}}$. The estimator $ACE(cc|x)_{\text{complete}}$ is generally preferred over $ACE(cc|x)_{\text{observed}}$. The main problem with the observed data parametric model (2.1) is that it does not require the fitted distribution obey the inequalities derived in Section 1.3 of Chapter 2. This indicates that the fitted multinomial regression model is not necessarily implied by Assumptions 2.1 - 2.7. One could develop novel parametrizations/link functions in order to incorporate covariates but always end up with distributions for the observables that obey the inequalities implied by the causal model. However this task would be complicated, and so we leave it to future research. Finally, there is a practical drawback to $ACE(cc|x)_{\text{observed}}$; the relationship between covariates and $ACE(cc|x)$ is transparent in the definition of $ACE(cc|x)_{\text{complete}}$ but unclear in definition of $ACE(cc|x)_{\text{observed}}$.

In future work, covariates could additionally be used to relax the exclusion restrictions in Assumptions 2.6 and 2.7. For example, in the one-level noncompliance setting (i.e. subjects may or may not comply with treatment assignment, but providers always comply with the treatment protocol), several methods have been proposed that relax exclusion restrictions. By using a Bayesian framework, models can be weakly identified without the exclusion restriction assumption, however the tradeoff is that they rely more heavily on auxiliary infor-

mation from proper priors and assumed parametric models [Imbens and Rubin, 1997, Hirano et al., 2000, Frangakis et al., 2002, Mattei et al., 2013]. In weakly identified models, ML estimates are not unique. Alternately, a ML approach could be used to establish identifiability by assuming functional relationships on covariates as opposed to the usual exclusion restrictions [Jo, 2002, Frangakis, 2006]. Either a Bayesian or ML approach to estimating causal estimands in the one-level noncompliance setting could be extended to the two-level noncompliance setting.

Chapter 3

ESTIMATING CONDITIONAL CAUSAL EFFECTS OF TREATMENT IN CLUSTERED RANDOMIZED CONTROLLED TRIALS

The methods proposed in Chapter 1 and 2 address both provider and subject noncompliance, however assume observations are independent and identically distributed (unconditionally or conditionally, respectively). In this chapter, we propose an estimator of the *average causal effect of treatment among provider-subject pairs that comply with assignment conditional on* x or $ACE(cc|x)$ that accounts for correlation within clusters defined by the providers. We extend the methods in Chapter 2 to the setting where treatment is randomized to subjects but clusters exist, as in the design of the motivating example considered in both Chapters 1 and 2. In Krupski et al. [2012], subjects were randomized in a 1:1 ratio to treatment or no treatment using permuted blocks stratified by clinic and factors known to affect outcomes. Subjects are naturally nested within providers.

In RCTs with clustering, outcomes and compliance behavior of the subjects of one provider are likely to resemble one another. Jo et al. [2008] address the possible impact of resemblance in subject compliance behavior in estimating the ITT effect, however they do not consider possible noncompliance of providers. Others have similarly considered clustering effects in settings where there is subject noncompliance, but do not additionally consider provider noncompliance [Frangakis et al., 2002, Albert, 2002, Brumback et al., 2013].

In this chapter, we define and propose estimators of a conditional causal effect of treatment in RCTs where both subject and provider noncompliance are present and subjects are nested within providers. For estimation, we take a Maximum Likelihood (ML) approach, extending the approaches of Chapter 2. We illustrate our proposal by re-examining the data

from Krupski et al. [2012].

3.1 Definition of the Conditional Causal Effect of Treatment

In Chapter 2 we motivated the need for defining and estimating a conditional causal effect of treatment. We now specifically address the setting where subjects are nested within providers, which is the case with our motivating example. Let $j = 1, ..., m$ index providers and $i = 1, ..., n_j$ index the subjects nested within provider j with $n = \sum_{j=1}^{m} n_j$. For the ith subject and jth provider, denote treatment assignment by Z_{ji} ($Z_{ji} = 1$ if the subject is assigned to treatment and $Z_{ji} = 0$ otherwise). The randomization does not influence the composition of provider-subject pairs, but only affects which provider-subject pairs get treatment. Denote the provider's observed compliance by D_{ji}^{P} and the subject's observed compliance by D_{ji}^{S}. If the subject was randomized to treatment ($Z_{ji} = 1$), $D_{ji}^{\mathrm{P}} = 1$ if the provider follows the treatment protocol and $D_{ji}^{\mathrm{P}} = 0$ otherwise and $D_{ji}^{\mathrm{S}} = 1$ if the subject takes the treatment and $D_{ji}^{\mathrm{S}} = 0$ otherwise. If the subject was randomized to no treatment ($Z_{ji} = 0$), then $D_{ji}^{\mathrm{P}} = D_{ji}^{\mathrm{S}} = 0$ since treatment was not available. Denote the subject's binary outcome by Y_{ji}. Let X_{ji}^{S} denote covariates collected on the subject and X_{j}^{P} denote covariates collected on the provider.

Let $\mathcal{N}$ denote the set of all n-dimensional column vectors of zeros and ones indicating all possible treatment assignments; hence $|\mathcal{N}| = 2^n$ and $\mathbf{z} \in \mathcal{N}$ represents one possible treatment assignment to the n provider-subject pairs. For the jith provider-subject pair, define the potential compliance under $\mathbf{z}$ by $D_{ji}^{\mathrm{P}}(\mathbf{z})$ and $D_{ji}^{\mathrm{S}}(\mathbf{z})$ for $\mathbf{z} \in \mathcal{N}$, and define the potential subject outcome under assignment $\mathbf{z}$ by $Y_{ji}(\mathbf{z}) \equiv Y_{ji}(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z}))$ for $\mathbf{z} \in \mathcal{N}$ where $\mathbf{D}^{\mathrm{P}}(\mathbf{z}) = (D_{11}^{\mathrm{P}}(\mathbf{z}), ..., D_{n_m m}^{\mathrm{P}}(\mathbf{z}))$ and $\mathbf{D}^{\mathrm{S}}(\mathbf{z}) = (D_{11}^{\mathrm{S}}(\mathbf{z}), ..., D_{n_m m}^{\mathrm{S}}(\mathbf{z}))$. There are 16 subgroups of provider-subject pairs defined by contrasting potential compliance under $z_{ji} = 1$ (treatment) and $z_{ji} = 0$ (no treatment) with $\mathbf{z}_{-ji}$ otherwise fixed. Denote the compliance type for the jith provider-subject pair by C_{ji}, which takes on values in $\{nn, cn, cc, nc\}$.

Define the conditional causal effect of treatment among provider-subject pairs of type cc

under assignment $\mathbf{z}$ for covariate value x as follows:

$$P(Y_{ji}(Z_{11} = z_1, ..., Z_{ji} = 1, ..., Z_{n_m,m} = z_{n_m,m}) = 1 \mid C_{ji} = cc, X_{ji} = x)$$
$$- P(Y_{ji}(Z_{11} = z_1, ..., Z_{ji} = 0, ..., Z_{n_m,m} = z_{n_m,m}) = 1 \mid C_{ji} = cc, X_{ji} = x),$$

In the following sections, we propose assumptions for defining a conditional causal estimand of interest as well as identify it.

3.2 A Conditional Causal Estimand of Interest: $ACE(cc|x)$

In Chapters 1 and 2, three assumptions are required to define a causal estimand of interest: Randomization, the Stable Unit Treatment Value Assumption (SUTVA), and Independent and Identical Distribution (iid) of the provider-subject pairs. In this chapter, we relax the iid assumption, extending the methods of Chapters 1 and 2 to the clustered RCT setting where provider-subject pairs are not all iid. Instead, we assume provider-subject pairs are only iid within clusters defined by provider.

First, subjects are randomized to MI intervention or no intervention, but the provider remains fixed.

Assumption 3.1 (Randomization) *For* $j = 1, ..., m$ *and* $i = 1, ..., n_j$,

$$Z_{ji} \perp\!\!\!\perp D^{\mathrm{P}}_{ji}(\mathbf{z}), D^{\mathrm{S}}_{ji}(\mathbf{z}), Y_{ji}(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})) \text{ for } \mathbf{z} \in \mathcal{N}.$$

Next, we suppose that there is no interference between subjects and that the observed data coincides with the potential values as follows:

Assumption 3.2 (SUTVA) *For* $\mathbf{z}$, $j = 1, ..., m$ *and* $i = 1, ..., n_j$,

$$D^{\mathrm{P}}_{ji}(\mathbf{z}) = D^{\mathrm{P}}_{ji}(z_{ji}),\ D^{\mathrm{S}}_{ji}(\mathbf{z}) = D^{\mathrm{S}}_{ji}(z_{ji}) \quad \text{and} \quad Y_{ji}(\mathbf{z}, \mathbf{D}^{\mathrm{P}}(\mathbf{z}), \mathbf{D}^{\mathrm{S}}(\mathbf{z})) = Y_{ji}(z_{ji}, D^{\mathrm{P}}_{ji}(z_{ji}), D^{\mathrm{S}}_{ji}(z_{ji})).$$

The observed data $(Y_{ji}, D^{\mathrm{P}}_{ji}, D^{\mathrm{S}}_{ji})$ *are related to the potential values as follows:*

$$D^{\mathrm{P}}_{ji} = \sum_{z \in \{0.1\}} D^{\mathrm{P}}_{ji}(z) \mathbb{1}(Z_{ji} = z),\ D^{\mathrm{S}}_{ji} = \sum_{z \in \{0,1\}} D^{\mathrm{S}}_{ji}(z) \mathbb{1}(Z_{ji} = z), \quad \text{and}$$

$$Y_{ji} = \sum_{z\in\{0,1\}} \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} Y_{ji}(z,u,v)\mathbb{1}(Z_{ji}=z, D^{\mathrm{P}}_{ji}(z)=u, D^{\mathrm{S}}_{ji}(z)=v).$$

The following assumption is where we diverge from Chapters 1 and 2. In this assumption, provider-subject pairs are assumed to be iid within clusters defined by provider.

Assumption 3.3 (iid within providers) *For $j = 1, ..., m$,*

$$Y_{ji} \mid Z_{ji}=z, C_{ji}=st, X_{ji}=x \overset{iid}{\sim} Bernoulli\big(\eta_{j,stzx}\big) \text{ for } st \in \{nn, cn, cc, nc\},\ z \in \{0,1\} \text{ and}$$

$$C_{ji} \mid X_{ji}=x \sim Multinomial_{16}\big(1, \pi = (\pi_{j,stx} : st \in \{nn, cn, cc, nc\})\big)$$

for $i = 1, ..., n_j$ and $x \in support(X) \equiv \mathcal{S}$.

Under Assumptions 3.1 - 3.3, we may define a provider-specific conditional causal estimand, denoted by $ACE(cc|x)_j$, defined as follows.

Definition 3.1 ($ACE(cc|x)_j$) *For the jth provider, define $ACE(cc|x)_j$ as*

$$\begin{aligned} ACE(cc|x)_j &= P(Y_{ji}=1 \mid C_{ji}=cc, Z_{ji}=1, X_{ji}=x) - P(Y_{ji}=1 \mid C_{ji}=cc, Z_{ji}=0, X_{ji}=x) \\ &= \eta_{j,cc1} - \eta_{j,cc0}. \end{aligned}$$

Similar to Chapters 1 and 2, under Assumptions 3.1 - 3.3 and the following three assumptions, it can be shown that $ACE(cc|x)_j$ is identifiable.

Assumption 3.4 (Compliance Type Restrictions) *For $j = 1, ..., m$ and $i = 1, ..., n_j$,*

$$\text{if } \ D^{\mathrm{P}}_{ji}(1) = D^{\mathrm{P}}_{ji}(0) \ \text{ then } \ D^{\mathrm{S}}_{ji}(1) = D^{\mathrm{S}}_{ji}(0).$$

Hence no provider-subject pairs have compliance type nc.

Assumption 3.5 (Stochastic Exclusion Restrictions)

$$\eta_{nn1x} = \eta_{nn0x} \equiv \eta_{stx} \text{ for } x \in \mathcal{S}.$$

Assumption 3.6 (Additional Stochastic Exclusion Restrictions)

$$\eta_{cn0x} = \eta_{nnx} \equiv \eta_{\bullet nx} \text{ for } x \in \mathcal{S}.$$

The "$\bullet$" in $\eta_{\bullet nx}$ denotes the provider-specific contribution to the compliance type may vary (e.g. first letter of the compliance type is c or n) but the subject-specific contribution does not (e.g. second letter is n).

Proposition 3.1 *Under Assumptions 3.1 - 3.6, $ACE(cc|x)_j$ is identifiable.*

The proof of Proposition 3.1 follows from the proof of Proposition 1.2 given in Appendix A, where each probability is replaced with a conditional probability on x specific to provider j.

Under Assumptions 3.1 - 3.4, $ACE(cc|x)_j$ is identifiable, but the number of parameters may be infinite and estimation is not possible. In order to estimate a causal estimand of interest, we assume a parametric model for the complete data.

Assumption 3.7 *Multilevel Model For a fixed vector of values x, we model $P(Y_{ji} = 1 \mid C_{ji} = st, Z_{ji} = z, X_{ji} = x)$ with a multilevel logistic regression model:*

$$\eta(x; \beta_\star, \epsilon_{j,\star}) = \frac{e^{\mathsf{X}_\eta \beta_\star + \mathsf{W}_\eta \epsilon_{j,\star}}}{1 + e^{\mathsf{X}_\eta \beta_\star + \mathsf{W}_\eta \epsilon_{j,\star}}} \text{ for } \star \in \{cc0, cc1, cn1, \bullet n\}$$

where X_η and W_η denote rows from the fixed and random effects design matrices that correspond to the values in x. We model $P(C_{ji} = st \mid X_{ji} = x) \equiv \pi(x, \zeta_{st}, \delta_{j,st})$ with a multilevel multinomial logistic regression model:

$$\pi(x, \zeta_{st}, \delta_{j,st}) = \frac{e^{\mathsf{X}_\pi \zeta_{st} + \mathsf{W}_\pi \delta_{j,st}}}{\sum_{st \in \{cc,nn,cn\}} e^{\mathsf{X}_\pi \zeta_{st} + \mathsf{W}_\pi \delta_{j,st}}} \text{ for } st \in \{cc, nn, cn\} \text{ with } \zeta_{cc} = \delta_{j,cc} = \mathbf{0}$$

where X_π and W_π denote rows from the fixed and random effects design matrices that correspond to the values in x. The random effects are assumed to be normally distributed:

$$\begin{pmatrix} \epsilon_j \\ \delta_j \end{pmatrix} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$$

where $\epsilon_j = (\epsilon_{j,cc0}, \epsilon_{j,cc1}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n})^\intercal$ and $\delta_j = (\delta_{j,cc}, \delta_{j,nn}, \delta_{j,cn})^\intercal$.

Under Assumptions 3.1 - 3.7, define the conditional causal estimand of interest, *the average conditional causal effect of treatment among provider-subject pairs that comply with assignment conditional on x*, denoted by $ACE(cc|x)$, as follows.

Definition 3.2 (*$ACE(cc|x)$*) *Let $\tilde{\jmath}$ be a therapist such that $\epsilon_{\tilde{\jmath}} = \delta_{\tilde{\jmath}} = 0$. In other words, $\tilde{\jmath}$ is interpreted as the average therapist in terms of compliance type and effect on patient outcomes. Define the average conditional causal effect of treatment among provider-subject pairs that comply with assignment conditional on x as*

$$\begin{aligned} ACE(cc|x) &= P(Y_{\tilde{\jmath}i} = 1 \mid C_{\tilde{\jmath}i} = cc, Z_{\tilde{\jmath}i} = 1, X_{\tilde{\jmath}i} = x) - \\ &\quad P(Y_{\tilde{\jmath}i} = 1 \mid C_{\tilde{\jmath}i} = cc, Z_{\tilde{\jmath}i} = 0, X_{\tilde{\jmath}i} = x) \\ &= \frac{e^{\mathsf{X}_\eta \beta_{cc1}}}{1 + e^{\mathsf{X}_\eta \beta_{cc1}}} - \frac{e^{\mathsf{X}_\eta \beta_{cc0}}}{1 + e^{\mathsf{X}_\eta \beta_{cc0}}}. \end{aligned}$$

The $ACE(cc|x)$ is interpreted as the average conditional causal effect of treatment among provider-subject pairs that comply with assignment conditional on x *for the average therapist*. The interpretation is similar to the conditional causal effect defined in Chapter 2, but critically, we allow for provider-specific effects whereas Chapter 2 does not. In the following sections, we propose a Maximum Likelihood (ML) approach to estimate $ACE(cc|x)$, $ACE(cc|x)_{\text{cluster}}$, and apply the estimator to data from a RCT of MI.

3.3 Proposed Estimator for $ACE(cc|x)$

Under Assumptions 3.1 - 3.7, there are then three possible compliance types: *cc*, *cn* and *nn*. The parameter of the model is (θ, Σ) where $\theta = (\xi, \zeta, \beta)$, $\zeta = (\zeta_{cn}, \zeta_{nn})$ and $\beta = (\beta_{cc0}, \beta_{cc1}, \beta_{cn1}, \beta_{\bullet n})$.

The complete data likelihood is

$$
\begin{aligned}
L \;=\; & \prod_{j=1}^{m}\left[\prod_{i=1}^{n_j}\Big[(1-\xi)\;\pi(X_{ji};\zeta_{nn},\delta_{j,nn})\Big]^{\mathbb{1}_{ji,nn000}}\;\big(1-\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,nn0000}}\;\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})^{\mathbb{1}_{ji,nn0001}}\;\times\right.\\
& \Big[\xi\;\pi(X_{ji};\zeta_{nn},\delta_{j,nn})\Big]^{\mathbb{1}_{ji,nn100}}\;\big(1-\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,nn1000}}\;\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})^{\mathbb{1}_{ji,nn1001}}\;\times\\
& \Big[(1-\xi)\;\pi(X_{ji};\zeta_{cn},\delta_{j,cn})\Big]^{\mathbb{1}_{ji,cn000}}\;\big(1-\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,cn0000}}\;\eta(X_{ji};\beta_{\bullet n},\epsilon_{j,\bullet n})^{\mathbb{1}_{ji,cn0001}}\;\times\\
& \Big[\xi\;\pi(X_{ji};\zeta_{cn},\delta_{j,cn})\Big]^{\mathbb{1}_{ji,cn110}}\;\big(1-\eta(X_{ji};\beta_{cn1},\epsilon_{j,cn1})\big)^{\mathbb{1}_{ji,cn1100}}\;\eta(X_{ji};\beta_{cn1},\epsilon_{j,cn1})^{\mathbb{1}_{ji,cn1101}}\;\times\\
& \Big[(1-\xi)\;\pi(X_{ji};\zeta_{cc},\delta_{j,cc})\Big]^{\mathbb{1}_{ji,cc000}}\;\big(1-\eta(X_{ji};\beta_{cc0},\epsilon_{j,cc0})\big)^{\mathbb{1}_{ji,cc0000}}\;\eta(X_{ji};\beta_{cc0},\epsilon_{j,cc0})^{\mathbb{1}_{ji,cc0001}}\;\times\\
& \left.\Big[\xi\;\pi(X_{ji};\zeta_{cc},\delta_{j,cc})\Big]^{\mathbb{1}_{ji,cc111}}\;\big(1-\eta(X_{ji};\beta_{cc1},\epsilon_{j,cc1})\big)^{\mathbb{1}_{ji,cc1110}}\;\eta(X_{ji};\beta_{cc1},\epsilon_{j,cc1})^{\mathbb{1}_{ji,cc1111}}\right]\;\times\\
& \phi\left(\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}\middle|\,\Sigma\right),
\end{aligned}
$$

where:

- $\phi(\cdot \mid \Sigma)$ is the density of a multivariate normal distribution with mean zero and covariance matrix Σ,

- $\mathbb{1}_{ji,stzuvy} = \mathbb{1}(C_{ji} = st, Z_{ji} = z, D_{ji}^{\mathrm{S}} = u, D_{ji}^{\mathrm{P}} = v, Y_{ji} = y)$ for $st \in \{nn, cn, cc\}$ and $z, u, v, y \in \{0, 1\}$,

- $\mathbb{1}_{ji,stzuv} = \mathbb{1}_{ji,stzuv0} + \mathbb{1}_{ji,stzuv1}$ for $st \in \{nn, cn, cc\}$ and $z, u, v \in \{0, 1\}$.

We use the MCEM-algorithm to estimate $ACE(cc|x)$. We treat the provider-specific random effects, $(\epsilon_{j,cc1}, \epsilon_{j,cc0}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n}, \delta_{j,nn}, \delta_{j,cn})$, as nuisance parameters, which are integrated

out in the E-step:

$$
\begin{aligned}
&E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) = \sum_{j=1}^{m} \int \left[\sum_{i=1}^{n_j} \Big\{ \big[\mathbb{1}(Z_{ji}=1)\log(\xi) + \mathbb{1}(Z_{ji}=0)\log(1-\xi) \big] + \right. \\
&(n^{(0)}_{ji,cn000} + n^{(0)}_{ji,cn110}) \log \pi(X_{ji}; \zeta_{cn}, \delta_{j,cn}) + (n^{(0)}_{ji,nn000} + n^{(0)}_{ji,nn100}) \log \pi(X_{ji}; \zeta_{nn}, \delta_{j,nn}) + \\
&n^{(0)}_{ji,cc0001} \log \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0}) + n^{(0)}_{ji,cc0000} \log(1 - \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0})) + \\
&n^{(0)}_{ji,cc1111} \log \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1}) + n^{(0)}_{ji,cc1110} \log(1 - \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1})) + \\
&n^{(0)}_{ji,cn1101} \log \eta(X_{ji}; \beta_{cn1}, \epsilon_{j,cn1}) + n^{(0)}_{ji,cn1100} \log(1 - \eta(X_{ji}; \beta_{cn1}, \epsilon_{j,cn1})) + \\
&(n^{(0)}_{ji,nn0001} + n^{(0)}_{ji,nn1001} + n^{(0)}_{ji,cn0001}) \log \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n}) + \\
&(n^{(0)}_{ji,nn0000} + n^{(0)}_{ji,nn1000} + n^{(0)}_{ji,cn0000}) \log(1 - \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n})) + \\
&\left. \log \phi\big((\epsilon_{j,cc1}, \epsilon_{j,cc0}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n}, \delta_{j,nn}, \delta_{j,cn})^{\intercal} \mid \Sigma\big) \right] \times \\
&f\big((\epsilon_{j,cc1}, \epsilon_{j,cc0}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n}, \delta_{j,nn}, \delta_{j,cn}) \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}, \theta^{(0)}, \Sigma^{(0)}\big) \times \\
&d\big((\epsilon_{j,cc1}, \epsilon_{j,cc0}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n}, \delta_{j,nn}, \delta_{j,cn})\big),
\end{aligned}
$$

where $n^{(0)}_{ji,stzuvy} \equiv P_{\theta^{(0)}}(C_{ji} = st, Z_{ji} = z, D^{\mathrm{P}}_{ji} = u, D^{\mathrm{S}}_{ji} = v, Y_{ji} = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ and $n^{(0)}_{ji,stzuv} = n^{(0)}_{ji,stzuv0} + n^{(0)}_{ji,stzuv1}$. We use Monte Carlo approximation to perform the integration [Wei and Tanner, 1990]. Details of the MCEM-algorithm are given in the Appendix C. Denote the ML estimate from the MCEM-algorithm by $\hat{\theta}$. The ML estimate for $ACE(cc|x)$, denoted by $ACE(cc|x)_{\text{cluster}}$ is

$$
ACE(cc|x)_{\text{cluster}} = \frac{e^{\mathsf{X}_\eta \hat{\beta}_{cc1}}}{1 + e^{\mathsf{X}_\eta \hat{\beta}_{cc1}}} - \frac{e^{\mathsf{X}_\eta \hat{\beta}_{cc0}}}{1 + e^{\mathsf{X}_\eta \hat{\beta}_{cc0}}}
$$

where X_η and W_η denote rows from the fixed and random effects design matrices that correspond to the values in x.

An estimate of the variance-covariance matrix of $\hat{\theta}$ can be estimated from the inverse of the observed data information matrix. The formula is given in the Appendix.

3.4 Application to a RCT of MI

We assume a model with one binary covariate, X_{ji}, which indicates the patient has a DAST-10 score ≥ 3. See Chapter 2 for more details. For fixed x, we model $P(Y_{ji} = 1 \mid C_{ji} =$

$st, Z_{ji} = z, X_{ji} = x) \equiv \eta(x; \beta_{stz}, \epsilon_j)$ with a multilevel logistic regression model and $P(C_{ji} = st \mid X_{ji} = x) \equiv \pi(x, \zeta_{st}, \delta_j)$ with a multilevel multinomial logistic regression model. The models are defined as follows:

$$\eta\left(x, \beta_\star = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}, \epsilon_{j,\star}\right) = \frac{e^{\beta_0+\beta_1 X_{ji}+\epsilon_{j,\star} Z_{ji}}}{1 + e^{\beta_0+\beta_1 X^1_{ji}+\epsilon_{j,\star} Z_{ji}}} \text{ for } \star \in \{cc0, cc1, cn1, \bullet n\}, \quad (3.1)$$

$$\pi\left(x, \zeta_{st} = \begin{pmatrix}\zeta_0 \\ \zeta_1\end{pmatrix}, \delta_{j,st}\right) = \frac{e^{\zeta_0+\zeta_1 X_{ji}+\delta_{j,st} Z_{ji}}}{\sum\limits_{st\in\{cc,cn,nn\}} e^{(1,X_{ji})\zeta_{st}+\delta_{j,st} Z_{ji}}} \text{ for } st \in \{nn, nc, cc\} \quad (3.2)$$

$$\text{with } \begin{pmatrix}\zeta_{cc} \\ \delta_{j,cc}\end{pmatrix} = \begin{pmatrix}0 \\ 0\end{pmatrix}.$$

The functional form of the model defined by (3.1) and (3.2) is similar to the model defined by (2.11) and (2.12) in Chapter 2, but includes provider-specific random effect terms. Since there are no therapists when $z = 0$ (i.e. we have a partially clustered design) we only assume random effects when $z = 1$ (see Baldwin et al. [2011]). The $ACE(cc|x)_{\text{complete}}$ from Chapter 2 was 0.05 (95% CI: [-0.32, 0.42]) and -0.21 (95% CI: [-1, 0.74]) for subjects with high and low drug use severity, respectively. The $ACE(cc|x)_{\text{cluster}}$ is -0.28 and -0.76 for subjects with high and low drug use severity, respectively.[1] The difference between $ACE(cc|x)_{\text{complete}}$ and $ACE(cc|x)_{\text{cluster}}$ may be due to within-cluster sample sizes were relatively small, which results in unstable results. The point estimates obtained from $ACE(cc|x)_{\text{cluster}}$ suggest a negative effect of MI on drug use, where the impact is more detrimental among subjects with low drug use severity.

[1]The formula for the 95% CI of these estimates is given in Appendix C, but have not been computed in this example.

3.5 Final Remarks

We extended the methods of Chapters 1 and 2 to address the clustered RCT design in the motivating example used in each proposal, a RCT of MI [Krupski et al., 2012]. Specifically, we extend the approach in Section 2.3.2 of Chapters 2 to account for clustering effects due to the fact that subjects are nested within providers.

The main challenge in incorporating the clustering effects lies in integrating out the random effect terms to obtain the marginal complete data log-likelihood $E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$, defined in Section 3.3. We took the approach of Booth and Hobert [1999] to perform the integration task, however other options include numerical quadrature methods or Gibbs sampling [McLachlan and Krishnan, 2007]. Regardless of the approach, approximating the integration causes considerable increases in computation time and resources. For this reason, we save further exploration into faster integration approximation and evaluation through simulation studies to future research.

Chapter 4

DISCUSSION

We proposed new statistical methodology for making causal inference in RCTs where provider and subject compliance underlie the definition of treatment receipt, yet provider and/or subject noncompliance is problematic. To begin, in Chapter 1 we assumed provider-subject pairs were iid and tackled the problem of estimating the marginal causal effect of treatment, $ACE(cc)$. We argued that $ACE(cc)$ can be identified under the assumptions outlined in Schochet and Chiang [2011], however these assumptions may not hold in some applications, specifically cognitive behavioral interventions. We proposed alternate assumptions for identification of $ACE(cc)$, which are more plausible in our motivating example, a RCT of MI. In Chapters 2 and 3 we extended the ML approach of Chapter 1 to tackle the problem of estimating the conditional causal effect of treatment, $ACE(cc|x)$, which are more relevant in the motivating example. In these chapters we take relaxed versions of the iid assumption of Chapter 1 to identify $ACE(cc|x)$, and proposed parametric modeling frameworks to estimate $ACE(cc|x)$.

In the motivating example data there were a total of 17 therapists and hence the assumption that provider-subject pairs are independent (i.e. Assumption 1.3 in Chapter 1 or Assumption 2.3 in Chapter 2) is strong. In addition, the cost of collecting compliance data is high because it requires humans to hand code transcriptions of the behavioral intervention. In the motivating example, due to the cost of collecting the observed compliance of providers and subjects, the example dataset is limited. Of the 868 recruited subjects, compliance data was only available for 57 of the 435 subjects assigned to $z = 1$. The validity of the results of the methods proposed is thus limited, as simulations suggest finite sample sizes must be much larger than the size of the motivating example dataset to obtain valid results. How-

ever, automating the collection of compliance data is an active area of research. In the near future, larger datasets are likely to be available, particularly among MI research, which will greatly increase the utility of our proposals.

The methods used in this research relied on Method of Moments and Maximum Likelihood approaches to estimate parameters of fully identified models. Future extensions of the research could relax identifying assumptions (Assumption 1.4-1.8 in Chapter 1, Assumptions 2.4-2.7 in Chapter 2 and Assumptions 3.4-3.6 in Chapter 3) by taking a Bayesian approach to estimation of $ACE(cc)$ and $ACE(cc|x)$.

BIBLIOGRAPHY

J.M. Albert. Estimating efficacy in clinical trials with clustered binary responses. *Statistics in Medicine*, 21:649–661, 2002.

J.D. Angrist and G.W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

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

S.A. Baldwin, D.J. Bauer, E. Stice, and P. Rohde. Evaluating models for partially clustered designs. *Psychological Methods*, 16(2):149–165, 2011.

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

J. Barnard, C.E. Frangakis, J.E. Hill, and D.B. Rubin. Principal stratification approach to broken randomized experiments: a case study of school choice voucher in New York City. *Journal of the American Statistical Association*, 98:299–311, 2003.

B. Bonet. Instrumentality tests revisited. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 48–55, 2001.

J.G. Booth and J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 16(1):265–285, 1999.

B.A. Brumback, He Z., M. Prasad, M.C. Freeman, and R. Rheingans. Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial. *Statistics in Medicine*, 33:1490–1502, 2013.

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

D.M. Chickering and J. Pearl. A clinician's tool for analyzing non-compliance. *Proceedings of the National Conference on Artificial Intelligence*, pages 1269–1276, 1996.

S.R. Cole and C.E. Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

J. Cuzick, R. Edward, and N. Segnan. Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*, 16:1017–1029, 1997.

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

P. Ding, Z. Geng, W. Wei Yan, and X.H. Zhou. Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106(496), 2012.

C. Dunn, D. Darnell, A. Carmel, D.C. Atkins, K. Bumgardner, and P. Roy-Byrne. Comparing the motivational interviewing integrity in two prevalent models of brief intervention service delivery for primary care settings. *Journal of Substance Abuse Treatment*, 51:47–52, 2015.

C. Frangakis and D. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379, 1999.

C. Frangakis, D. Rubin, and X.H. Zhou. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and applications to advance directive forms. *Biostatistics*, 3(2):147–164, 2002.

C.E. Frangakis. Comment on causal effects in the presence of non compliance: a latent variable interpretation, by A. Forcina. *Metron*, LXIV:8–14, 2006.

C.E. Frangakis and D.B. Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.

J. Geweke. *Handbook of Computational Economics*, chapter 15. Elsevier, Amsterdam, North-Holland, 1996.

P.B. Gilbert, R.J. Bosch, and M.G. Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, 59:531–541, 2003.

D.F. Heitjan. Ignorability and bias in clinical trials. *Statistics in Medicine*, 18:2421–2434, 1999.

M.A. Hernan and J.M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17:360–372, 2006.

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

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

N.S. Ialongo, L. Werthamer, S.G. Kellam, C.H. Brown, S. Wang, and Y. Lin. Proximal impact of two first-grade preventative interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology*, 27:599–642, 1999.

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

B. Jo. Estimation of intervention effects with noncompliance: alternative model specifications. *Journal of Educational and Behavioral Statistics*, 27:385–420, 2002.

B. Jo, T. Asparouhov, and B.O. Muthen. Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine*, 27:5565–5577, 2008.

K.M. Johnston, P. Gustafson, A.R. Levy, and P. Grootendorst. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27: 1539–1556, 2008.

A. Krupski, J.M. Joesch, C. Dunn, D. Donovan, K. Bumgardner, S.P. Lord, R. Ries, and P. Roy-Byrne. Testing the effects of brief intervention in primary care for problem drug use in a randomized controlled trial: rationale, design, and methods. *Addiction Science and Clinical Practice*, 7(1):1–10, 2012.

M.J. Lambert. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. Wiley, New Jersey, 5th edition, 2013.

K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 57(2):425–437, 1995.

D.E. Levy, J.A. O'Malley, and S.L. Normand. Covariate adjustment in clinical trials with non-ignorable missing data and non-compliance. *Statistics in Medicine*, 23:2319–2339, 2004.

T.A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44(1):226–232, 1982.

Y. Ma, J. Roy, and B. Marcus. Causal models for randomized trials with two active treatments and continuous compliance. *Statistics in Medicine*, 30:2349–2362, 2011.

A. Mattei, F. Mealli, and B. Pacini. *Advances in Theoretical and Applied Statistics*, chapter 22. Springer, New York, 2013.

C. McDonald, S. Hiu, and W. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing*, 9:304–312, 1992.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

R.M. Miller and S. Rollnick. The effectiveness and ineffectiveness of complex behavioral interventions: Impact of treatment fidelity. *Contemporary Clinical Trials*, 37:234241, 2014.

W.R. Miller and R.S. Rose. Toward a theory of motivational interviewing. *American Psychology*, 64(6):527–537, 2010.

T.B. Moyers, T. Martin, D. Catley, K.J. Harris, and J.S. Ahluwalia. Assessing the integrity of motivational interventions: reliability of the motivational interviewing skills code. *Behavioral and Cognitive Psychotherapy*, 31:177–184, 2003.

T.B. Moyers, J.K. Manuel, S.M. Hendrickson, and W.R. Miller. Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28(1): 19–26, 2005.

N. Nagelkerke, V. Fidler, R. Bersen, and M. Borgdorff. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19: 1849–1864, 2000.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472, 1990.

J.A. O'Malley and S.L. Normand. Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics*, 61:325–334, 2005.

J. Pearl. Aspects of graphical models connected with causality. *Technical Report R-195-LL, Computer Science Department, UCLA*, 1993.

J. Pearl. On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pages 435–443, 1995.

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

Y. Peng, R. Little, and T.E. Raghuanthan. An extended general location model for causal inference from data subject to noncompliance and missing values. *Biometrics*, 60:598–607, 2004.

R.R. Ramsahai and S.L. Lauritzen. Likelihood analysis of the binary instrumental variable model. *Biometrika*, 98(4):987–994, 2011.

T.S. Richardson and J.M. Robins. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, chapter 25. College Publications London, UK, 2010.

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

JM Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.

D.B. Rubin. Estimating causal effects of treatment in randomized and non-randomized studies. *Journal of Education and Psychology*, 66:688–701, 1974.

D.B. Rubin. Bayesian inference for causal effects. *Annals of Statistics*, 6:34–58, 1978.

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

D.B. Rubin. Statistics and causal inference: comment: which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1980b.

P.Z. Schochet and H.S. Chiang. Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Education and Behavioral Statistics*, 36(3):307–345, 2011.

S. Schwartz, Gatto N.M., and U.B. Campbell. Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA). *Epidemiologic Perspectives and Innovations*, 9(3):1–11, 2012.

B.E. Shepherd, P.B. Gilbert, Y. Jemiai, and A. Rotnitzky. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*, 62(2):332–342, 2006.

A. Sommer and S.L. Zeger. On estimating efficacy from clinical trials. *Statistics in Medicine*, 10:45–52, 1991.

L. Taylor and X.H. Zhou. Relaxing latent ignorability in the ITT analysis of randomized studies with missing data and noncompliance. *UW Biostatistics Working Paper Series*, (257), 2009a.

L. Taylor and X.H. Zhou. Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *UW Biostatistics Working Paper Series*, (312), 2009b.

T.J. VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B*, 65:817–835, 2003.

G.C.G. Wei and M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

L.H.Y. Yau and R.J. Little. Inference for 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*, 96: 1232–1244, 2001.

D.F. Zatzick, D.M. Donovan, C. Dunn, G.J. Jurkovich, J. Wang, J. Russo, F.P. Rivara, C.D. Zatzick, J.R. Love, C.R. McFadden, and L.M. Gentilello. Disseminating Organizational Screening and Brief Intervention Services (DO-SBIS) for alcohol at trauma centers study design. *General Hospital Psychiatry*, 35:174–180, 2013.

X.H. Zhou and S.M. Li. ITT analysis of randomized encouragement design studies with missing data. *Statistics in Medicine*, 25:2737–2761, 2006.

Appendix A

APPENDIX FOR CHAPTER 1

A.1 Proof of Proposition 1.2

For $st \in \mathcal{C}$, and $z, u, v \in \{0, 1\}$, define the following quantities:

$$\begin{aligned} \pi_{st} &= P(C_i = st), \\ \psi_{stzuv} &= P(C_i = st \mid Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v), \\ \eta_{stz} &= P(Y_i = 1 \mid C_i = st, Z_i = z). \end{aligned}$$

By Assumptions 1.4 and 1.5, there are 7 compliance types: $\mathcal{C}_7 = \{cc, cn, ca, nn, na, an, aa\}$. So $\pi_{st} = \psi_{stzuv} = \eta_{stz} = 0$ for $st \in \mathcal{C} \setminus \mathcal{C}_7$.

First consider π_{na}.

$$\begin{aligned} \pi_{na} &= P(C_i = na \mid Z_i = 0)P(Z_i = 0) + P(C_i = na \mid Z_i = 1)P(Z_i = 1) \\ &= P(C_i = na \mid Z_i = 1) \text{ by Assumptions 1.1 and 1.2} \\ &= P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1 \mid Z_i = 1). \end{aligned}$$

Similarly, $\pi_{an} = P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 0)$. And by Assumption 1.5, $\pi_{nn} = P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)$, and $\pi_{aa} = P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 0)$.

Next consider conditional therapist-patient compliance type, conditioned on the observed data, denoted by ψ_{stzuv} for $st \in \mathcal{C}_7$ and $z, u, v \in \{1, 0\}$. Because Assumptions 1.4 and 1.5 eliminate certain compliance types,

$$\psi_{an010} = \psi_{nn100} = \psi_{aa011} = \psi_{na101} = 1.$$

Consider ψ_{nn000}.

$$\begin{aligned}
\psi_{nn000} &= P(C_i = nn \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&= P(C_i = nn, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) \\
&= P(C_i = nn \mid Z_i = 0)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) \\
&= P(C_i = nn \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) \text{ by Assumptions 1.1 and 1.2} \\
&= P(C_i = nn, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0) \\
&= P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0).
\end{aligned}$$

By using similar arguments, the following can also be shown:

$$\begin{aligned}
\psi_{na001} &= P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1 \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1 \mid Z_i = 0), \\
\psi_{aa111} &= P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 0)/P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 1), \\
\psi_{an110} &= P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 0)/P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 1), \\
\psi_{ca111} &= \psi_{ca001} P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1 \mid Z_i = 0)/P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 1) \text{ and} \\
\psi_{cn000} &= \psi_{cn110} P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 0).
\end{aligned}$$

By the identities $1 = \psi_{na001} + \psi_{ca001}$ and $1 = \psi_{an110} + \psi_{cn110}$, ψ_{ca001}, ψ_{ca111}, ψ_{cn110} and ψ_{cn000} are identified. By the identities $1 = \psi_{nn000} + \psi_{cn000} + \psi_{cc000}$ and $1 = \psi_{aa111} + \psi_{ca111} + \psi_{cc111}$, ψ_{cc000} and ψ_{cc111} are identified.

Now consider π_{cn}, π_{ca} and π_{cc}. These probabilities can be written as follows:

$$\begin{aligned}
\pi_{cn} &= \psi_{cn110} P(Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0) + \psi_{cn000} P(Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0), \\
\pi_{ca} &= \psi_{ca001} P(Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 1) + \psi_{ca111} P(Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1), \\
\pi_{cc} &= \psi_{cc000} P(Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) + \psi_{cc111} P(Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1).
\end{aligned}$$

Hence π_{cn}, π_{ca} and π_{cc} are identified.

Finally, consider the parameters η_{cc0} and η_{cc1}.

$$\begin{aligned}
\eta_{cc0} &= P(Y_i = 1 \mid C_i = cc, Z_i = 0) \\
&= P(Y_i = 1 \mid C_i = cc, Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&= P(Y_i = 1, C_i = cc \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0)/P(C_i = cc \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&= P(Y_i = 1, C_i = cc \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0)/\psi_{cc000}.
\end{aligned}$$

By Assumptions 1.4 and 1.5, there are only three compliance types that correspond to the observed triple $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}) = (0, 0, 0)$, compliance types cc, cn and nn. So by the Law of Total Probability,

$$\begin{aligned}
&P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&\quad = \sum_{st \in \{cc,cn,nn\}} P(Y_i = 1, C_i = st \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&\quad = P(Y_i = 1, C_i = cc \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) + \\
&\qquad P(Y_i = 1 \mid C_i = nn, Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0)(\psi_{cn000} + \psi_{nn000}) \text{ by Assumption 1.8.}
\end{aligned}$$

Consider $P(Y_i = 1 \mid C_i = nn, Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0)$. This quantity can be written as

$$\begin{aligned}
&P(Y_i = 1 \mid C_i = nn, Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \\
&\quad = P(Y_i = 1 \mid C_i = nn, Z_i = 1, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) \text{ by Assumption 1.6} \\
&\quad = P(Y_i = 1, C_i = nn, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)/P(C_i = nn, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1) \\
&\quad = P(Y_i = 1, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)/P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1) \text{ by Assumptions 1.4 and 1.5.}
\end{aligned}$$

Hence η_{cc0} can be identified.

$$\eta_{cc0} = \frac{1}{\psi_{cc000}}\left(P(Y_i = 1 \mid Z_i = 0, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0) - (\psi_{cn000} + \psi_{nn000})\frac{P(Y_i = 1, D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)}{P(D_i^{\mathrm{P}} = 0, D_i^{\mathrm{S}} = 0 \mid Z_i = 1)}\right).$$

Similarly, we can identify η_{cc1}.

$$\eta_{cc1} = \frac{1}{\psi_{cc111}}\left(P(Y_i = 1 \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1) - (\psi_{ca111} + \psi_{aa111})\frac{P(Y_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 0)}{P(D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 1 \mid Z_i = 0)}\right).$$

The $ACE(cc) = \eta_{cc1} - \eta_{cc0}$ is therefore identified.

Note that the observed data for the ith provider-subject pair is comprised of four binary variables $(Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i)$ and hence can be represented by a 16-vector $(\mathbf{n}_i)$ with a Multinomial distribution:

$$\mathbf{n}_i \sim Multinomial_{16}(1, \boldsymbol{\rho})$$

where elements of $\mathbf{n}_i$ and $\boldsymbol{\rho}$ are indexed by $z, u, v, y \in \{0, 1\}$. Furthermore, $\rho_{zuvy} = P(Z_i = z, D^{\mathrm{P}} = u, D^{\mathrm{S}} = v, Y_i = y)$, and we can equivalently write η_{cc0} and η_{cc1} as follows:

$$\begin{aligned}
\eta_{cc0} &= \frac{(\rho_{0001}\rho_{100\bullet} + \rho_{1001}\rho_{010\bullet})\rho_{1\bullet\bullet\bullet} - (\rho_{1001}\rho_{100\bullet} + \rho_{1001}\rho_{110\bullet})\rho_{0\bullet\bullet\bullet}}{(\rho_{000\bullet}\rho_{100\bullet} + \rho_{100\bullet}\rho_{010\bullet})\rho_{1\bullet\bullet\bullet} - (\rho_{100\bullet}\rho_{100\bullet} + \rho_{100\bullet}\rho_{110\bullet})\rho_{0\bullet\bullet\bullet}}, \\
\eta_{cc1} &= \frac{(\rho_{1111}\rho_{011\bullet} + \rho_{0111}\rho_{101\bullet})\rho_{0\bullet\bullet\bullet} - (\rho_{0111}\rho_{011\bullet} + \rho_{0111}\rho_{001\bullet})\rho_{1\bullet\bullet\bullet}}{(\rho_{111\bullet}\rho_{011\bullet} + \rho_{011\bullet}\rho_{101\bullet})\rho_{0\bullet\bullet\bullet} - (\rho_{011\bullet}\rho_{011\bullet} + \rho_{011\bullet}\rho_{001\bullet})\rho_{1\bullet\bullet\bullet}},
\end{aligned}$$

where $\rho_{zuv\bullet} = \sum_{y=0,1} \rho_{zuvy}$ and $\rho_{z\bullet\bullet\bullet} = \sum_{u=0,1} \sum_{v=0,1} \sum_{y=0,1} \rho_{zuvy}$.

A.2 EM Algorithm for the Obtaining $ACE(cc)_{\mathrm{ML}}$

The parameters of the model are $\theta = (\eta, \pi, \xi)$. $\eta = (\eta_{cc0}, \eta_{cc1}, \eta_{cn1}, \eta_{ca0}, \eta_{an}, \eta_{na}, \eta_{\bullet n}, \eta_{\bullet a})$ where $\eta_{stz} = P(Y_i = 1 \mid Z_i = z, C_i = st)$. $\pi = (\pi_{nn}, \pi_{na}, \pi_{an}, \pi_{aa}, \pi_{cn}, \pi_{ca}, \pi_{cc})$ where $\pi_{st} = P(C_i = st)$. $\xi = P(Z_i = 1)$.

A.2.1 E-step

For the E-step of the EM-algorithm, choose initial values for θ, and denote them by $\theta^{(0)}$. The conditional expectation of the complete data log-likelihood given the observed data under the

initial parameter values is:

$$
\begin{aligned}
&E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}) = \sum_{i=1}^{n} \Big\{ \mathbb{1}(Z_i = 1) \log(\xi) + \mathbb{1}(Z_i = 0) \log(1-\xi) + \\
&(n^{(0)}_{i,aa011} + n^{(0)}_{i,aa111}) \log \pi_{aa} + (n^{(0)}_{i,na001} + n^{(0)}_{i,na101}) \log \pi_{na} + (n^{(0)}_{i,an010} + n^{(0)}_{i,an110}) \log \pi_{an} + \\
&(n^{(0)}_{i,nn000} + n^{(0)}_{i,nn100}) \log \pi_{nn} + (n^{(0)}_{i,cn000} + n^{(0)}_{i,cn110}) \log \pi_{cn} + (n^{(0)}_{i,ca001} + n^{(0)}_{i,ca111}) \log \pi_{ca} + \\
&n^{(0)}_{i,cc0001} \log \eta_{cc0} + n^{(0)}_{i,cc0000} \log(1-\eta_{cc0}) + n^{(0)}_{i,cc1111} \log \eta_{cc1} + n^{(0)}_{i,cc1110} \log(1-\eta_{cc1}) + \\
&n^{(0)}_{i,cn1101} \log \eta_{cn1} + n^{(0)}_{i,cn1100} \log(1-\eta_{cn1}) + n^{(0)}_{i,ca0011} \log \eta_{ca0} + n^{(0)}_{i,ca0010} \log(1-\eta_{ca0}) + \\
&(n^{(0)}_{i,nn0001} + n^{(0)}_{i,nn1001} + n^{(0)}_{i,cn0001}) \log \eta_{\bullet n} + (n^{(0)}_{i,nn0000} + n^{(0)}_{i,nn1000} + n^{(0)}_{i,cn0000}) \log(1-\eta_{\bullet n}) + \\
&(n^{(0)}_{i,aa0111} + n^{(0)}_{i,aa1111} + n^{(0)}_{i,ca1111}) \log \eta_{\bullet a} + (n^{(0)}_{i,aa0110} + n^{(0)}_{i,aa1110} + n^{(0)}_{i,ca1110}) \log(1-\eta_{\bullet a}) + \\
&(n^{(0)}_{i,an0101} + n^{(0)}_{i,an1101}) \log \eta_{an} + (n^{(0)}_{i,an0100} + n^{(0)}_{i,an1100}) \log(1-\eta_{an}) + \\
&(n^{(0)}_{i,na0011} + n^{(0)}_{i,na1011}) \log \eta_{na} + (n^{(0)}_{i,na0010} + n^{(0)}_{ji,na1010}) \log(1-\eta_{na}) \Big\}
\end{aligned}
$$

where $n^{(0)}_{i,stzuvy} \equiv P_{\theta^{(0)}}(C_i = st, Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$ and $n^{(0)}_{i,stzuv} = n^{(0)}_{i,stzuv0} + n^{(0)}_{i,stzuv1}$. As an example, consider $n^{(0)}_{i,an110y}$ for $y \in \{0,1\}$.

$$
\begin{aligned}
n^{(0)}_{i,an110y} &= P_{\theta^{(0)}}(C_i = an, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}) \\
&= P_{\theta^{(0)}}(C_i = an, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y) \\
&= \mathbb{1}(Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y) P_{\theta^{(0)}}(C_i = an \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y)
\end{aligned}
$$

and by Baye's rule,

$$
\begin{aligned}
&P_{\theta^{(0)}}(C_i = an \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y) \\
&\quad = \frac{P_{\theta^{(0)}}(Y_i = y, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid C_i = an) P_{\theta^{(0)}}(C_i = an)}{P_{\theta^{(0)}}(Y_i = y, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0)} \\
&\quad = \frac{P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = an) P_{\theta^{(0)}}(Z_i = 1) P_{\theta^{(0)}}(C_i = an)}{\sum_{st=\{an,cn\}} P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = st) P_{\theta^{(0)}}(Z_i = 1) P_{\theta^{(0)}}(C_i = st)} \\
&\quad = \frac{\big(\eta^{(0)}_{an}\big)^y \big(1-\eta^{(0)}_{an}\big)^{1-y} \pi^{(0)}_{an}}{\big(\eta^{(0)}_{an}\big)^y \big(1-\eta^{(0)}_{an}\big)^{1-y} \pi^{(0)}_{an} + \big(\eta^{(0)}_{cn1}\big)^y \big(1-\eta^{(0)}_{cn1}\big)^{1-y} \pi^{(0)}_{cn}}.
\end{aligned}
$$

The values of $n^{(0)}_{stzuvy}$ for $st \in \mathcal{C}_7$ and $z, u, v, y \in \{0,1\}$ are similarly obtained.

A.2.2 M-step

In the M-step of the EM-algorithm, we set $\frac{\partial}{\partial\theta}E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$ equal to zero, solve for θ and denote the solutions by $\theta^{(1)}$. First consider $\frac{\partial}{\partial\eta_{\bullet n}}E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$.

$$\frac{\partial E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})}{\partial\eta_{\bullet n}} = \sum_{i=1}^{n}\left(\frac{n^{(0)}_{i,nn0001} + n^{(0)}_{i,cn0001} + n^{(0)}_{i,nn1001}}{\eta_{\bullet n}} - \frac{n^{(0)}_{i,nn0000} + n^{(0)}_{i,cn0000} + n^{(0)}_{i,nn1000}}{1-\eta_{\bullet n}}\right).$$

Setting the right-hand side of the above equation to zero and rearranging terms, we have:

$$\eta^{(1)}_{\bullet n} = \frac{\sum_{i=1}^{n}\left[n^{(0)}_{i,nn0001} + n^{(0)}_{i,cn0001} + n^{(0)}_{i,nn1001}\right]}{\sum_{i=1}^{n}\left[n^{(0)}_{i,nn000} + n^{(0)}_{i,cn000} + n^{(0)}_{i,nn100}\right]}$$

where $n^{(0)}_{i,stzuv} = n^{(0)}_{i,stzuv0} + n^{(0)}_{i,stzuv1}$. The updates $\eta^{(1)}_{\bullet a}$, $\eta^{(1)}_{ca0}$, $\eta^{(1)}_{cn1}$, $\eta^{(1)}_{cc0}$, $\eta^{(1)}_{cc1}$, $\eta^{(1)}_{an}$ and $\eta^{(1)}_{na}$ are obtained similarly.

Next consider $\frac{\partial}{\partial\pi_{st}}E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$ for $st \in \mathcal{C}_7 = \{\pi_{nn}, \pi_{na}, \pi_{an}, \pi_{aa}, \pi_{cn}, \pi_{ca}, \pi_{cc}\}$. Note that $\sum_{st\in\mathcal{C}_7}\pi_{st} = 1$, so to incorporate this constraint let $\pi_{cc} = 1 - \sum_{st\in\mathcal{C}_7\setminus\{cc\}}\pi_{st}$. First consider $\frac{\partial}{\partial\pi_{nn}}E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$.

$$\frac{\partial E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})}{\partial\pi_{nn}} = \sum_{i=1}^{n}\left[\frac{n^{(0)}_{i,nn000} + n^{(0)}_{i,nn100}}{\pi_{nn}} - \frac{n^{(0)}_{i,cc000} + n^{(0)}_{i,cc111}}{\pi_{cc}}\right].$$

Setting the right-hand side of the above equation to zero and rearranging terms, we have:

$$\pi^{(1)}_{nn} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,nn000} + n^{(0)}_{i,nn100}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000} + n^{(0)}_{i,cc111}]}.$$

Similarly,

$$\pi^{(1)}_{aa} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,aa011}+n^{(0)}_{i,aa111}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000}+n^{(0)}_{i,cc111}]}, \quad \pi^{(1)}_{an} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,an010}+n^{(0)}_{i,an110}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000}+n^{(0)}_{i,cc111}]}, \quad \pi^{(1)}_{na} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,na001}+n^{(0)}_{i,na101}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000}+n^{(0)}_{i,cc111}]},$$

$$\pi^{(1)}_{ca} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,ca001}+n^{(0)}_{i,ca111}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000}+n^{(0)}_{i,cc111}]}, \quad \pi^{(1)}_{cn} = \pi^{(1)}_{cc}\frac{\sum_{i=1}^{n}[n^{(0)}_{i,cn000}+n^{(0)}_{i,cn110}]}{\sum_{i=1}^{n}[n^{(0)}_{i,cc000}+n^{(0)}_{i,cc111}]}.$$

Taking the sum of the six previous equations, we have

$$1 - \pi_{cc}^{(1)} = \pi_{cc}^{(1)} \frac{n - \sum_{i=1}^{n} [n_{i,cc000}^{(0)} + n_{i,cc111}^{(0)}]}{\sum_{i=1}^{n} [n_{i,cc000}^{(0)} + n_{i,cc111}^{(0)}]}.$$

Hence $\pi_{cc}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} [n_{i,cc000}^{(0)} + n_{i,cc111}^{(0)}]$, and the remaining updates for π simplify to the following:

$$\pi_{nn}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,nn000}^{(0)} + n_{i,nn100}^{(0)}], \quad \pi_{aa}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,aa011}^{(0)} + n_{i,aa111}^{(0)}],$$

$$\pi_{an}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,an010}^{(0)} + n_{i,an110}^{(0)}], \quad \pi_{na}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,na001}^{(0)} + n_{i,na101}^{(0)}],$$

$$\pi_{ca}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,ca001}^{(0)} + n_{i,ca111}^{(0)}], \quad \pi_{cn}^{(1)} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [n_{i,cn000}^{(0)} + n_{i,cn110}^{(0)}].$$

Finally, consider $\frac{\partial}{\partial \xi} E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})$.

$$\frac{\partial E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y})}{\partial \xi} = \sum_{i=1}^{n} \left[\frac{\mathbb{1}(Z_i = 1)}{\xi} - \frac{\mathbb{1}(Z_i = 0)}{1 - \xi} \right].$$

Setting the right-hand side of the above equation to zero and rearranging terms, we have

$$\xi^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Z_i = 1).$$

We have now obtained $\theta^{(1)}$, and the E- and M-steps are iterated until the updates for θ converge.

Since the EM algorithm only guarantees a local maxima is reached, we repeat the algorithm at multiple starting values. We try many different initial values $\theta^{(0)}$ and choose the solution that corresponds to the highest converged likelihood value, which we denote by $\hat{\theta}$. The ML estimator for $ACE(cc)$ is defined as $ACE(cc)_{\mathrm{ML}} = \hat{\eta}_{cc1} - \hat{\eta}_{cc0}$.

A.2.3 Variance of the $ACE(cc)_{\mathrm{ML}}$

To obtain an estimate for the variance of $ACE(cc)_{\mathrm{ML}}$, we first must obtain the observed data information matrix. Since computing the gradient or Hessian of the observed data likelihood is cumbersome, we employ Louis's formula for Multinomial data [Louis, 1982]. Suppose the parameter of the model is $\theta = (\eta, \pi, \xi)$ where $\eta = (\eta_{cc0}, \eta_{cc1}, \eta_{cn1}, \eta_{ca0}, \eta_{an}, \eta_{na}, \eta_{\bullet n}, \eta_{\bullet a})$ and $\pi =$

$(\pi_{cn}, \pi_{ca}, \pi_{an}, \pi_{na}, \pi_{nn}, \pi_{aa})$. For convenience we will write $\pi_{cc} \equiv 1-\pi_{nn}-\pi_{aa}-\pi_{an}-\pi_{na}-\pi_{ca}-\pi_{cn}$. For one provider-subject pair, the observed data has a Multinomial distribution:

$$\mathbf{n}_i \sim Multinomial_{16}(1, \boldsymbol{\rho} \equiv \boldsymbol{\rho}(\theta))$$

and the complete data also has a Multinomial distribution:

$$\mathbf{n}_i^* \sim Multinomial_{28}(1, \boldsymbol{\rho}^* \equiv \boldsymbol{\rho}^*(\theta)),$$

for $i = 1, ..., n$. The elements of $\mathbf{n}_i$ and $\boldsymbol{\rho}$ are indexed by $z, u, v, y \in \{0, 1\}$ and the elements of $\mathbf{n}_i^*$ and $\boldsymbol{\rho}^*$ are indexed by $st \in \mathcal{C}_7 = \{cc, cn, ca, an, na, nn, aa\}$ and $z, y \in \{0, 1\}$. The probability vectors $\boldsymbol{\rho}$ and $\boldsymbol{\rho}^*$ are defined by the following tables.

st	$\boldsymbol{\rho}^*_{st00}$	$\boldsymbol{\rho}^*_{st01}$	$\boldsymbol{\rho}^*_{st10}$	$\boldsymbol{\rho}^*_{st11}$
cc	$(1-\xi)\pi_{cc}(1-\eta_{cc0})$	$(1-\xi)\pi_{cc}\eta_{cc0}$	$\xi\pi_{cc}(1-\eta_{cc1})$	$\xi\pi_{cc}\eta_{cc1}$
cn	$(1-\xi)\pi_{cn}(1-\eta_{\bullet n})$	$(1-\xi)\pi_{cn}\eta_{\bullet n}$	$\xi\pi_{cn}(1-\eta_{cn1})$	$\xi\pi_{cn}\eta_{cn1}$
ca	$(1-\xi)\pi_{ca}(1-\eta_{ca0})$	$(1-\xi)\pi_{cn}\eta_{ca0}$	$\xi\pi_{ca}(1-\eta_{\bullet a})$	$\xi\pi_{ca}\eta_{\bullet a}$
an	$(1-\xi)\pi_{an}(1-\eta_{an})$	$(1-\xi)\pi_{an}\eta_{an}$	$\xi\pi_{an}(1-\eta_{an})$	$\xi\pi_{an}\eta_{an}$
na	$(1-\xi)\pi_{na}(1-\eta_{na})$	$(1-\xi)\pi_{na}\eta_{na}$	$\xi\pi_{na}(1-\eta_{na})$	$\xi\pi_{na}\eta_{na}$
nn	$(1-\xi)\pi_{nn}(1-\eta_{\bullet n})$	$(1-\xi)\pi_{nn}\eta_{\bullet n}$	$\xi\pi_{nn}(1-\eta_{\bullet n})$	$\xi\pi_{nn}\eta_{\bullet n}$
aa	$(1-\xi)\pi_{aa}(1-\eta_{\bullet a})$	$(1-\xi)\pi_{aa}\eta_{\bullet a}$	$\xi\pi_{aa}(1-\eta_{\bullet a})$	$\xi\pi_{aa}\eta_{\bullet a}$

z	y	$\boldsymbol{\rho}_{z00y}$	$\boldsymbol{\rho}_{z01y}$	$\boldsymbol{\rho}_{z10y}$	$\boldsymbol{\rho}_{z11y}$
0	0	$\boldsymbol{\rho}^*_{cc00} + \boldsymbol{\rho}^*_{cn00} + \boldsymbol{\rho}^*_{nn00}$	$\boldsymbol{\rho}^*_{ca00} + \boldsymbol{\rho}^*_{na00}$	$\boldsymbol{\rho}^*_{an00}$	$\boldsymbol{\rho}^*_{aa00}$
0	1	$\boldsymbol{\rho}^*_{cc01} + \boldsymbol{\rho}^*_{cn01} + \boldsymbol{\rho}^*_{nn01}$	$\boldsymbol{\rho}^*_{ca01} + \boldsymbol{\rho}^*_{na01}$	$\boldsymbol{\rho}^*_{an01}$	$\boldsymbol{\rho}^*_{aa01}$
1	0	$\boldsymbol{\rho}^*_{nn10}$	$\boldsymbol{\rho}^*_{na10}$	$\boldsymbol{\rho}^*_{cn10} + \boldsymbol{\rho}^*_{an10}$	$\boldsymbol{\rho}^*_{cc10} + \boldsymbol{\rho}^*_{ca10} + \boldsymbol{\rho}^*_{aa10}$
1	1	$\boldsymbol{\rho}^*_{nn11}$	$\boldsymbol{\rho}^*_{na11}$	$\boldsymbol{\rho}^*_{cn11} + \boldsymbol{\rho}^*_{an11}$	$\boldsymbol{\rho}^*_{cc11} + \boldsymbol{\rho}^*_{ca11} + \boldsymbol{\rho}^*_{aa11}$

In other words, the following relationships hold:

$$
\begin{array}{llll}
n_{i,0000} = n^*_{i,cc00} + n^*_{i,cn00} + n^*_{i,nn00}, & n_{i,1100} = n^*_{i,cn10} + n^*_{i,an10}, & n_{i,1010} = n^*_{i,na10}, & n_{i,0110} = n^*_{i,aa00}, \\
n_{i,0001} = n^*_{i,cc01} + n^*_{i,cn01} + n^*_{i,nn01}, & n_{i,1101} = n^*_{i,cn11} + n^*_{i,an11}, & n_{i,1011} = n^*_{i,na11}, & n_{i,0111} = n^*_{i,aa01}, \\
n_{i,1110} = n^*_{i,cc10} + n^*_{i,ca10} + n^*_{i,aa10}, & n_{i,0010} = n^*_{i,ca00} + n^*_{i,na00}, & n_{i,0100} = n^*_{i,an00}, & n_{i,1000} = n^*_{i,nn10}, \\
n_{i,1111} = n^*_{i,cc11} + n^*_{i,ca11} + n^*_{i,aa11}, & n_{i,0011} = n^*_{i,ca01} + n^*_{i,na01}, & n_{i,0101} = n^*_{i,an01}, & n_{i,1001} = n^*_{i,nn11}.
\end{array}
$$

The complete data log-likelihood for one observation, denoted by $\lambda(\mathbf{n}^*_i, \theta)$, is:

$$
\begin{aligned}
\lambda(\mathbf{n}^*_i, \theta) = {} & n^*_{i,cc00}\big(\log(1-\xi) + \log(\pi_{cc}) + \log(1-\eta_{cc0})\big) + n^*_{i,cc01}\big(\log(1-\xi) + \log(\pi_{cc}) + \log(\eta_{cc0})\big) + \\
& n^*_{i,cc10}\big(\log(\xi) + \log(\pi_{cc}) + \log(1-\eta_{cc1})\big) + n^*_{i,cc11}\big(\log(\xi) + \log(\pi_{cc}) + \log(\eta_{cc1})\big) + \\
& n^*_{i,cn00}\big(\log(1-\xi) + \log(\pi_{cn}) + \log(1-\eta_{\bullet n})\big) + n^*_{i,cn01}\big(\log(1-\xi) + \log(\pi_{cn}) + \log(\eta_{\bullet n})\big) + \\
& n^*_{i,cn10}\big(\log(\xi) + \log(\pi_{cn}) + \log(1-\eta_{cn1})\big) + n^*_{i,cn11}\big(\log(\xi) + \log(\pi_{cn}) + \log(\eta_{cn1})\big) + \\
& n^*_{i,ca00}\big(\log(1-\xi) + \log(\pi_{ca}) + \log(1-\eta_{ca0})\big) + n^*_{i,ca01}\big(\log(1-\xi) + \log(\pi_{ca}) + \log(\eta_{ca0})\big) + \\
& n^*_{i,ca10}\big(\log(\xi) + \log(\pi_{ca}) + \log(1-\eta_{\bullet a})\big) + n^*_{i,ca11}\big(\log(\xi) + \log(\pi_{ca}) + \log(\eta_{\bullet a})\big) + \\
& n^*_{i,an00}\big(\log(1-\xi) + \log(\pi_{an}) + \log(1-\eta_{an})\big) + n^*_{i,an01}\big(\log(1-\xi) + \log(\pi_{an}) + \log(\eta_{an})\big) + \\
& n^*_{i,an10}\big(\log(\xi) + \log(\pi_{an}) + \log(1-\eta_{an})\big) + n^*_{i,an11}\big(\log(\xi) + \log(\pi_{an}) + \log(\eta_{an})\big) + \\
& n^*_{i,na00}\big(\log(1-\xi) + \log(\pi_{na}) + \log(1-\eta_{na})\big) + n^*_{i,na01}\big(\log(1-\xi) + \log(\pi_{na}) + \log(\eta_{na})\big) + \\
& n^*_{i,na10}\big(\log(\xi) + \log(\pi_{na}) + \log(1-\eta_{na})\big) + n^*_{i,na11}\big(\log(\xi) + \log(\pi_{na}) + \log(\eta_{na})\big) + \\
& n^*_{i,nn00}\big(\log(1-\xi) + \log(\pi_{nn}) + \log(1-\eta_{\bullet n})\big) + n^*_{i,nn01}\big(\log(1-\xi) + \log(\pi_{nn}) + \log(\eta_{\bullet n})\big) + \\
& n^*_{i,nn10}\big(\log(\xi) + \log(\pi_{nn}) + \log(1-\eta_{\bullet n})\big) + n^*_{i,nn11}\big(\log(\xi) + \log(\pi_{nn}) + \log(\eta_{\bullet n})\big) + \\
& n^*_{i,aa00}\big(\log(1-\xi) + \log(\pi_{aa}) + \log(1-\eta_{\bullet a})\big) + n^*_{i,aa01}\big(\log(1-\xi) + \log(\pi_{aa}) + \log(\eta_{\bullet a})\big) + \\
& n^*_{i,aa10}\big(\log(\xi) + \log(\pi_{aa}) + \log(1-\eta_{\bullet a})\big) + n^*_{i,aa11}\big(\log(\xi) + \log(\pi_{aa}) + \log(\eta_{\bullet a})\big).
\end{aligned}
$$

The gradient vector of $\lambda(\mathbf{n}_i^*, \theta)$, denoted by $S(\mathbf{n}_i^*, \theta)$, is:

$$
\begin{aligned}
S(\mathbf{n}_i^*, \theta) = \Bigg(& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{cc0}} = \frac{n^*_{i,cc01}}{\eta_{cc0}} - \frac{n^*_{i,cc00}}{1-\eta_{cc0}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{cc1}} = \frac{n^*_{i,cc11}}{\eta_{cc1}} - \frac{n^*_{i,cc10}}{1-\eta_{cc1}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{cn1}} = \frac{n^*_{i,cn11}}{\eta_{cn1}} - \frac{n^*_{i,cn10}}{1-\eta_{cn1}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{ca0}} = \frac{n^*_{i,ca01}}{\eta_{ca0}} - \frac{n^*_{i,ca00}}{1-\eta_{ca0}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{an}} = \frac{n^*_{i,an11} + n^*_{i,an01}}{\eta_{an}} - \frac{n^*_{i,an10} + n^*_{i,an00}}{1-\eta_{an}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{na}} = \frac{n^*_{i,na11} + n^*_{i,na01}}{\eta_{na}} - \frac{n^*_{i,na10} + n^*_{i,na00}}{1-\eta_{na}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{\bullet n}} = \frac{n^*_{i,nn11} + n^*_{i,nn01} + n^*_{i,cn01}}{\eta_{\bullet n}} - \frac{n^*_{i,nn10} + n^*_{i,nn00} + n^*_{i,cn00}}{1-\eta_{\bullet n}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \eta_{\bullet a}} = \frac{n^*_{i,aa11} + n^*_{i,aa01} + n^*_{i,ca11}}{\eta_{\bullet a}} - \frac{n^*_{i,aa10} + n^*_{i,aa00} + n^*_{i,ca10}}{1-\eta_{\bullet a}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{cn}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cnzy}}{\pi_{cn}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{ca}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cazy}}{\pi_{ca}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{an}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,anzy}}{\pi_{an}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{na}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,nazy}}{\pi_{na}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{cc}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,nnzy}}{\pi_{nn}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{aa}} = \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,aazy}}{\pi_{aa}} - \frac{\sum_{z=0,1}\sum_{y=0,1} n^*_{i,cczy}}{\pi_{cc}}, \\
& \frac{\partial \lambda(\mathbf{n}_i^*, \theta)}{\partial \pi_{cn}} = \frac{\sum_{st\in\mathcal{C}_7}(n^*_{i,st10} + n^*_{i,st11})}{\xi} - \frac{\sum_{st\in\mathcal{C}_7}(n^*_{i,st00} + n^*_{i,st01})}{1-\xi} \Bigg)^T.
\end{aligned}
$$

By Louis's formula, the observed data Fisher information matrix, $I_{\mathbf{n}}$, can be computed from the EM algorithm as:

$$I_{\mathbf{n}} = \sum_{i=1}^{n} S(\hat{\mathbf{n}}_i^*, \hat{\theta}) S(\hat{\mathbf{n}}_i^*, \hat{\theta})^T$$

where $\hat{\mathbf{n}}_i^* = E_{\hat{\theta}}(\mathbf{n}_i^* \mid \mathbf{n}_i)$ and $\hat{\theta}$ is the ML estimate of θ obtained from the EM algorithm. By the relationships above,

$$\hat{n}^*_{i,cc00} = \frac{n_{i,0000}(1-\hat{\eta}_{cc0})\hat{\pi}_{cc}}{(1-\hat{\eta}_{cc0})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{cn}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{nn}}, \qquad \hat{n}^*_{i,ca00} = \frac{n_{i,0010}(1-\hat{\eta}_{ca0})\hat{\pi}_{ca}}{(1-\hat{\eta}_{ca0})\hat{\pi}_{ca}+(1-\hat{\eta}_{na})\hat{\pi}_{na}},$$

$$\hat{n}^*_{i,cc01} = \frac{n_{i,0001}\hat{\eta}_{cc0}\hat{\pi}_{cc}}{\hat{\eta}_{cc0}\hat{\pi}_{cc}+\hat{\eta}_{\bullet n}\hat{\pi}_{cn}+\hat{\eta}_{\bullet n}\hat{\pi}_{nn}}, \qquad \hat{n}^*_{i,ca01} = \frac{n_{i,0011}\hat{\eta}_{ca0}\hat{\pi}_{ca}}{\hat{\eta}_{ca0}\hat{\pi}_{ca}+\hat{\eta}_{na}\hat{\pi}_{na}},$$

$$\hat{n}^*_{i,cc10} = \frac{n_{i,1110}(1-\hat{\eta}_{cc1})\hat{\pi}_{cc}}{(1-\hat{\eta}_{cc1})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{ca}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{aa}}, \qquad \hat{n}^*_{i,ca10} = \frac{n_{i,1110}(1-\hat{\eta}_{\bullet a})\hat{\pi}_{ca}}{(1-\hat{\eta}_{cc1})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{ca}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{aa}},$$

$$\hat{n}^*_{i,cc11} = \frac{n_{i,1111}\hat{\eta}_{cc1}\hat{\pi}_{cc}}{\hat{\eta}_{cc1}\hat{\pi}_{cc}+\hat{\eta}_{\bullet a}\hat{\pi}_{ca}+\hat{\eta}_{\bullet a}\hat{\pi}_{aa}}, \qquad \hat{n}^*_{i,ca11} = \frac{n_{i,1111}\hat{\eta}_{\bullet a}\hat{\pi}_{ca}}{\hat{\eta}_{cc1}\hat{\pi}_{cc}+\hat{\eta}_{\bullet a}\hat{\pi}_{ca}+\hat{\eta}_{\bullet a}\hat{\pi}_{aa}},$$

$$\hat{n}^*_{i,cn00} = \frac{n_{i,0000}(1-\hat{\eta}_{\bullet n})\hat{\pi}_{cn}}{(1-\hat{\eta}_{cc0})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{cn}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{nn}}, \qquad \hat{n}^*_{i,nn00} = \frac{n_{i,0000}(1-\hat{\eta}_{\bullet n})\hat{\pi}_{nn}}{(1-\hat{\eta}_{cc0})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{cn}+(1-\hat{\eta}_{\bullet n})\hat{\pi}_{nn}},$$

$$\hat{n}^*_{i,cn01} = \frac{n_{i,0001}\hat{\eta}_{\bullet n}\hat{\pi}_{cn}}{\hat{\eta}_{cc0}\hat{\pi}_{cc}+\hat{\eta}_{\bullet n}\hat{\pi}_{cn}+\hat{\eta}_{\bullet n}\hat{\pi}_{nn}}, \qquad \hat{n}^*_{i,nn01} = \frac{n_{i,0001}\hat{\eta}_{\bullet n}\hat{\pi}_{nn}}{\hat{\eta}_{cc0}\hat{\pi}_{cc}+\hat{\eta}_{\bullet n}\hat{\pi}_{cn}+\hat{\eta}_{\bullet n}\hat{\pi}_{nn}},$$

$$\hat{n}^*_{i,cn10} = \frac{n_{i,1100}(1-\hat{\eta}_{cn1})\hat{\pi}_{cn}}{(1-\hat{\eta}_{cn1})\hat{\pi}_{cn}+(1-\hat{\eta}_{an})\hat{\pi}_{an}}, \qquad \hat{n}^*_{i,nn10} = n_{i,1000},$$

$$\hat{n}^*_{i,cn11} = \frac{n_{i,1101}\hat{\eta}_{cn1}\hat{\pi}_{cn}}{\hat{\eta}_{cn1}\hat{\pi}_{cn}+\hat{\eta}_{an}\hat{\pi}_{an}}, \qquad \hat{n}^*_{i,nn11} = n_{i,1001},$$

$$\hat{n}^*_{i,an00} = n_{i,0100}, \qquad \hat{n}^*_{i,aa00} = n_{i,0110},$$

$$\hat{n}^*_{i,an01} = n_{i,0101}, \qquad \hat{n}^*_{i,aa01} = n_{i,0111},$$

$$\hat{n}^*_{i,an10} = \frac{n_{i,1100}(1-\hat{\eta}_{an})\hat{\pi}_{an}}{(1-\hat{\eta}_{cn1})\hat{\pi}_{cn}+(1-\hat{\eta}_{an})\hat{\pi}_{an}}, \qquad \hat{n}^*_{i,aa10} = \frac{n_{i,1110}(1-\hat{\eta}_{\bullet a})\hat{\pi}_{aa}}{(1-\hat{\eta}_{cc1})\hat{\pi}_{cc}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{ca}+(1-\hat{\eta}_{\bullet a})\hat{\pi}_{aa}},$$

$$\hat{n}^*_{i,an11} = \frac{n_{i,1101}\hat{\eta}_{an}\hat{\pi}_{an}}{\hat{\eta}_{cn1}\hat{\pi}_{cn}+\hat{\eta}_{an}\hat{\pi}_{an}}, \qquad \hat{n}^*_{i,aa11} = \frac{n_{i,1111}\hat{\eta}_{\bullet a}\hat{\pi}_{aa}}{\hat{\eta}_{cc1}\hat{\pi}_{cc}+\hat{\eta}_{\bullet a}\hat{\pi}_{ca}+\hat{\eta}_{\bullet a}\hat{\pi}_{aa}},$$

$$\hat{n}^*_{i,na00} = \frac{n_{i,0010}(1-\hat{\eta}_{na})\hat{\pi}_{na}}{(1-\hat{\eta}_{ca0})\hat{\pi}_{ca}+(1-\hat{\eta}_{na})\hat{\pi}_{na}}, \qquad \hat{n}^*_{i,na10} = n_{i,1010},$$

$$\hat{n}^*_{i,na01} = \frac{n_{i,0011}\hat{\eta}_{na}\hat{\pi}_{na}}{\hat{\eta}_{ca0}\hat{\pi}_{ca}+\hat{\eta}_{na}\hat{\pi}_{na}}, \qquad \hat{n}^*_{i,na11} = n_{i,1011}.$$

An asymptotically consistent estimate of the variance of $ACE(cc)_{\mathrm{ML}}$ is then

$$\widehat{Var}(ACE(cc)_{\mathrm{ML}}) = (-1, 1, 0, ..., 0) I_{\mathbf{n}}^{-1} (-1, 1, 0, ..., 0)^T.$$

Appendix B

APPENDIX FOR CHAPTER 2

B.1 The EM Gradient Algorithm

The complete data log-likelihood for one observation is:

$$
\begin{aligned}
\ell(C_i, Z_i, D_i, Y_i, X_i, \theta) = {} & Z_i \log(\xi) + (1 - Z_i) \log(1 - \xi) + \\
& \mathbb{1}_{i,nn000}\big[Y_i \log \eta(X_i; \beta_{\bullet n}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet n})) + \log \pi(X_i; \zeta_{nn})\big] + \\
& \mathbb{1}_{i,nn100}\big[Y_i \log \eta(X_i; \beta_{\bullet n}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet n})) + \log \pi(X_i; \zeta_{nn})\big] + \\
& \mathbb{1}_{i,cn000}\big[Y_i \log \eta(X_i; \beta_{\bullet n}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet n})) + \log \pi(X_i; \zeta_{cn})\big] + \\
& \mathbb{1}_{i,cn110}\big[Y_i \log \eta(X_i; \beta_{cn1}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{cn1})) + \log \pi(X_i; \zeta_{cn})\big] + \\
& \mathbb{1}_{i,an010}\big[Y_i \log \eta(X_i; \beta_{an}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{an})) + \log \pi(X_i; \zeta_{an})\big] + \\
& \mathbb{1}_{i,an110}\big[Y_i \log \eta(X_i; \beta_{an}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{an})) + \log \pi(X_i; \zeta_{an})\big] + \\
& \mathbb{1}_{i,na001}\big[Y_i \log \eta(X_i; \beta_{na}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{na})) + \log \pi(X_i; \zeta_{na})\big] + \\
& \mathbb{1}_{i,na101}\big[Y_i \log \eta(X_i; \beta_{na}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{na})) + \log \pi(X_i; \zeta_{na})\big] + \\
& \mathbb{1}_{i,aa011}\big[Y_i \log \eta(X_i; \beta_{\bullet a}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet a})) + \log \pi(X_i; \zeta_{aa})\big] + \\
& \mathbb{1}_{i,aa111}\big[Y_i \log \eta(X_i; \beta_{\bullet a}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet a})) + \log \pi(X_i; \zeta_{aa})\big] + \\
& \mathbb{1}_{i,ca111}\big[Y_i \log \eta(X_i; \beta_{\bullet a}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{\bullet a})) + \log \pi(X_i; \zeta_{ca})\big] + \\
& \mathbb{1}_{i,ca001}\big[Y_i \log \eta(X_i; \beta_{ca0}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{ca0})) + \log \pi(X_i; \zeta_{ca})\big] + \\
& \mathbb{1}_{i,cc111}\big[Y_i \log \eta(X_i; \beta_{cc1}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{cc1})) + \log \pi(X_i; \zeta_{cc})\big] + \\
& \mathbb{1}_{i,cc000}\big[Y_i \log \eta(X_i; \beta_{cc0}) + (1 - Y_i) \log(1 - \eta(X_i; \beta_{cc0})) + \log \pi(X_i; \zeta_{cc})\big],
\end{aligned}
$$

where $\mathbb{1}_{i,stzuv} = \mathbb{1}(C_i = st, Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v)$.

B.1.1 E-step

For the E-step, choose initial values for θ, and denote them by $\theta^{(0)}$. The conditional expectation of the joint complete data log-likelihood given the observed data under the initial parameter values is

$$
\begin{aligned}
E_{\theta^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \Big\{ & \big[\mathbb{1}(Z_i = 1)\log(\xi) + \mathbb{1}(Z_i = 0)\log(1-\xi)\big] + \\
& (n^{(0)}_{i,aa011} + n^{(0)}_{i,aa111}) \log \pi(X_i, \zeta_{aa}, \delta_{j,aa}) + (n^{(0)}_{i,na001} + n^{(0)}_{i,na101}) \log \pi(X_i, \zeta_{na}, \delta_{j,na}) + \\
& (n^{(0)}_{i,an010} + n^{(0)}_{i,an110}) \log \pi(X_i, \zeta_{an}, \delta_{j,an}) + (n^{(0)}_{i,nn000} + n^{(0)}_{i,nn100}) \log \pi(X_i, \zeta_{nn}, \delta_{j,nn}) + \\
& (n^{(0)}_{i,cn000} + n^{(0)}_{i,cn110}) \log \pi(X_i, \zeta_{cn}, \delta_{j,cn}) + (n^{(0)}_{i,ca001} + n^{(0)}_{i,ca111}) \log \pi(X_i, \zeta_{ca}, \delta_{j,ca}) + \\
& n^{(0)}_{i,cc0001} \log \eta(X_i, \beta_{cc0}, \epsilon_{j,cc0}) + n^{(0)}_{i,cc0000} \log(1 - \eta(X_i, \beta_{cc0}, \epsilon_{j,cc0})) + \\
& n^{(0)}_{i,cc1111} \log \eta(X_i, \beta_{cc1}, \epsilon_{j,cc1}) + n^{(0)}_{i,cc1110} \log(1 - \eta(X_i, \beta_{cc1}, \epsilon_{j,cc1})) + \\
& n^{(0)}_{i,cn1101} \log \eta(X_i, \beta_{cn1}, \epsilon_{j,cn1}) + n^{(0)}_{i,cn1100} \log(1 - \eta(X_i, \beta_{cn1}, \epsilon_{j,cn1})) + \\
& n^{(0)}_{i,ca0011} \log \eta(X_i, \beta_{ca0}, \epsilon_{j,ca0}) + n^{(0)}_{i,ca0010} \log(1 - \eta(X_i, \beta_{ca0}, \epsilon_{j,ca0})) + \\
& (n^{(0)}_{i,nn0001} + n^{(0)}_{i,nn1001} + n^{(0)}_{i,cn0001}) \log \eta(X_i, \beta_{\bullet n}, \epsilon_{j,\bullet n}) + \\
& (n^{(0)}_{i,nn0000} + n^{(0)}_{i,nn1000} + n^{(0)}_{i,cn0000}) \log(1 - \eta(X_i, \beta_{\bullet n}, \epsilon_{j,\bullet n})) + \\
& (n^{(0)}_{i,aa0111} + n^{(0)}_{i,aa1111} + n^{(0)}_{i,ca1111}) \log \eta(X_i, \beta_{\bullet a}, \epsilon_{j,\bullet a}) + \\
& (n^{(0)}_{i,aa0110} + n^{(0)}_{i,aa1110} + n^{(0)}_{i,ca1110}) \log(1 - \eta(X_i, \beta_{\bullet a}, \epsilon_{j,\bullet a})) + \\
& (n^{(0)}_{i,an0101} + n^{(0)}_{i,an1101}) \log \eta(X_i, \beta_{an}, \epsilon_{j,an}) + (n^{(0)}_{i,an0100} + n^{(0)}_{i,an1100}) \log(1 - \eta(X_i, \beta_{an}, \epsilon_{j,an})) + \\
& (n^{(0)}_{i,na0011} + n^{(0)}_{i,na1011}) \log \eta(X_i, \beta_{na}, \epsilon_{j,na}) + (n^{(0)}_{i,na0010} + n^{(0)}_{i,na1010}) \log(1 - \eta(X_{ji}; \beta_{na}, \epsilon_{j,na})) \Big\},
\end{aligned}
$$

where $n^{(0)}_{i,stzuvy} \equiv P_{\theta^{(0)}}(C_i = st, Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ and $n^{(0)}_{i,stzuv} = n^{(0)}_{i,stzuv0} + n^{(0)}_{i,stzuv1}$. As an example, consider $n^{(0)}_{i,an110y}$.

$$
\begin{aligned}
n^{(0)}_{i,an110y} &= P_{\theta^{(0)}}(C_i = st, Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \\
&= P_{\theta^{(0)}}(C_i = an, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y, X_i) \\
&= \mathbb{1}(Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y) P_{\theta^{(0)}}(C_i = an \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y, X_i).
\end{aligned}
$$

By Baye's rule,

$$\begin{aligned}
&P_{\theta^{(0)}}(C_i = an \mid Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0, Y_i = y, X_i) \\
&\quad = \frac{P_{\theta^{(0)}}(Y_i = y, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid C_i = an, X_i) P_{\theta^{(0)}}(C_i = an \mid X_i)}{P_{\theta^{(0)}}(Y_i = y, Z_i = 1, D_i^{\mathrm{P}} = 1, D_i^{\mathrm{S}} = 0 \mid X_i)} \\
&\quad = \frac{P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = an, X_i) P_{\theta^{(0)}}(C_i = an \mid X_i)}{\sum_{st=\{an,cn\}} P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = st, X_i) P_{\theta^{(0)}}(C_i = st \mid X_i)}
\end{aligned}$$

where

$$\begin{aligned}
P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = an, X_i) &= \eta(X_i; \beta_{an}^{(0)})^y \big(1 - \eta(X_i; \beta_{an}^{(0)})\big)^{1-y}, \\
P_{\theta^{(0)}}(Y_i = y \mid Z_i = 1, C_i = cn, X_i) &= \eta(X_i; \beta_{cn1}^{(0)})^y \big(1 - \eta(X_i; \beta_{cn1}^{(0)})\big)^{1-y}, \\
P_{\theta^{(0)}}(C_i = an \mid X_i) &= \pi(X_i; \zeta_{an}^{(0)}), \\
P_{\theta^{(0)}}(C_i = cn \mid X_i) &= \pi(X_i; \zeta_{cn}^{(0)}).
\end{aligned}$$

The values of $n_{i,stzuvy}^{(0)}$ for $st \in \mathcal{C}_7$ and $z, u, v \in \{0, 1\}$ are similarly obtained.

Note that $n_{i,anzuvy}^{(0)} = 0$ for $(u, v) \neq (1, 0)$ and that these quantities do not appear in $E_{\theta^{(0)}}(\ell | \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$. For brevity, from this point forward we let $n_{i,an1}^{(0)} \equiv \sum_{u,v\in\{0,1\}} n_{i,an1uv}^{(0)} = n_{i,an110}^{(0)}$ and $n_{i,an0}^{(0)} \equiv \sum_{u,v\in\{0,1\}} n_{i,an0uv}^{(0)} = n_{i,an010}^{(0)}$. For all $st \in \mathcal{C}_7$ and $z \in \{0, 1\}$ we define $n_{i,stz}^{(0)}$ similarly.

B.1.2 M-step

In the M-step, we obtain the update $\theta^{(k+1)}$ as follows:

$$\theta^{(k+1)} = \theta^{(k)} + \left(-\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}} E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \Big|_{\theta=\theta^{(k)}} \right)^{-1} \left(\frac{\partial}{\partial\theta} E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \Big|_{\theta=\theta^{(k)}} \right).$$

The score of the conditional expectation of the joint complete data log-likelihood evaluated at $\theta^{(k)}$, $\frac{\partial}{\partial\theta}E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})|_{\theta=\theta^{(k)}}$, is:

$$\sum_{i=1}^{n} \begin{pmatrix} \frac{Z_i}{\xi^{(k)}} - \frac{1-Z_i}{1-\xi^{(k)}} \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(n_{i,aa0}^{(k)} + n_{i,aa1}^{(k)} - \pi(X_i; \zeta_{aa}^{(k)})\right) \\ \vdots \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(n_{i,ca0}^{(k)} + n_{i,ca1}^{(k)} - \pi(X_i; \zeta_{ca}^{(k)})\right) \\ \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} n_{i,cc0}^{(k)}\left(Y_i - \eta(X_i; \beta_{cc0}^{(k)})\right) \\ \vdots \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(n_{i,na0}^{(k)} + n_{i,na1}^{(k)}\right)\left(Y_i - \eta(X_i; \beta_{na}^{(k)})\right) \end{pmatrix}.$$

The negative Hessian of the conditional expectation of the joint complete data log-likelihood evaluated at $\theta^{(k)}$ is given by the following block diagonal matrix:

$$\frac{\partial^2}{\partial\theta\partial\theta^\intercal} E_{\theta^{(k)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})\Big|_{\theta=\theta^{(k)}} =$$

$$\sum_{i=1}^{n} \left(\begin{array}{c|c|c} \frac{Z_i}{(\xi^{(k)})^2} + \frac{1-Z_i}{(1-\xi^{(k)})^2} & 0 & 0 \\ \hline 0 & \begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \left(\pi(X_i;\zeta_{aa}^{(k)}) - \pi(X_i;\zeta_{aa}^{(k)})^2\right) & 0 \\ & -\begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \pi(X_i;\zeta_{na}^{(k)})\pi(X_i;\zeta_{aa}^{(k)}) \quad \ddots & \\ & \vdots & \\ \hline 0 & 0 & \begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal n_{i,cc0}^{(k)}\left(\eta(X_i;\beta_{cc0}^{(k)}) - \eta(X_i;\beta_{cc0}^{(k)})^2\right) \\ & & 0 \quad \ddots \\ & & \vdots \end{array}\right).$$

The E- and M-steps are repeated until the updates converge. Since the EM algorithm only guarantees a local maxima is reached, we repeat the algorithm at multiple starting values. We try many different initial values $\theta^{(0)}$ and choose the solution that corresponds to the highest converged likelihood value, which we denote by $\hat{\theta}$. The ML estimate for $ACE(cc|x)$ for a fixed value of x is

$$ACE(cc|x)_{\text{complete}} = \eta(x; \hat{\beta}_{cc1}) - \eta(x; \hat{\beta}_{cc0}).$$

B.1.3 Variance

The asymptotic variance of $ACE(cc|x)_{\text{complete}}$ is obtained from the inverse observed data information matrix, $I_{\mathbf{n}}$. Since computing the score or Hessian of the observed data likelihood is extremely cumbersome, we employ Louis's formula [Louis, 1982]:

$$\begin{aligned}
I_{\mathbf{n}} &= E_{\hat{\theta}}\Big(- \frac{\partial^2(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial\theta\partial\theta^{\intercal}}\Big|_{\theta=\hat{\theta}}\Big) - Cov_{\hat{\theta}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big) \\
&= \sum_{i=1}^{n} E_{\hat{\theta}}\Big(- \frac{\partial^2(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial\theta\partial\theta^{\intercal}}\Big|_{\theta=\hat{\theta}}\Big) \qquad \text{(B.1)} \\
&\quad - \sum_{i=1}^{n} E_{\hat{\theta}}\Big(\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial\theta}\Big|_{\theta=\hat{\theta}} \frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial\theta}\Big|_{\theta=\hat{\theta}}^{\intercal}\Big) \qquad \text{(B.2)} \\
&\quad + \sum_{i=1}^{n} E_{\hat{\theta}}\Big(\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial\theta}\Big|_{\theta=\hat{\theta}}\Big) E_{\hat{\theta}}\Big(\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial\theta}\Big|_{\theta=\hat{\theta}}^{\intercal}\Big) \qquad \text{(B.3)} \\
&\equiv \sum_{i=1}^{n} \Big[E_{\hat{\theta}}\big(- H_i(\hat{\theta})\big) - E_{\hat{\theta}}\big(S_i(\hat{\theta}) S_i(\hat{\theta})^{\intercal}\big) + E_{\hat{\theta}}\big(S_i(\hat{\theta})\big) E_{\hat{\theta}}\big(S_i(\hat{\theta})\big)^{\intercal}\Big]
\end{aligned}$$

The second equality holds due to the assumed independence of observations. Note that in the summands of (B.1) and (B.3), differentiation and expectation can be interchanged due to the fact that the likelihood is a linear function of $\mathbb{1}(C_i)$. Hence we have already computated (B.1) and (B.3) in Section B.1.2.

Consider the summands in (B.2). First, $S_i(\hat{\theta})$ is defined as follows:

$$S_i(\hat{\theta}) = \begin{pmatrix} S_i(\hat{\xi}) \\ \hline S_i(\hat{\zeta}) \\ \hline S_i(\hat{\beta} \end{pmatrix} = \begin{pmatrix} \left(\frac{Z_i}{\hat{\xi}} - \frac{1-Z_i}{1-\hat{\xi}}\right) \\ \hline \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(\mathbb{1}_{i,aa011} + \mathbb{1}_{i,aa111} - \pi(X_i; \hat{\zeta}_{aa})\right) \\ \vdots \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(\mathbb{1}_{i,ca001} + \mathbb{1}_{i,ca111} - \pi(X_i; \hat{\zeta}_{ca})\right) \\ \hline \begin{pmatrix} 1 \\ X_i \end{pmatrix} \mathbb{1}_{i,cc000}\big(Y_i - \eta(X_i; \hat{\beta}_{cc0})\big) \\ \vdots \\ \begin{pmatrix} 1 \\ X_i \end{pmatrix} \big(\mathbb{1}_{i,na001} + \mathbb{1}_{i,na101}\big)\big(Y_i - \eta(X_i; \hat{\beta}_{na})\big) \end{pmatrix}.$$

Since $E_{\hat{\theta}}(\mathbb{1}_{i,stzuv}\mathbb{1}_{i,s't'z'u'v'} \mid Z_i = z, D_i^{\mathrm{P}} = u, D^{\mathrm{S}} = v, Y_i, X_i) = 0$ for $st \neq s't'$, $E_{\hat{\theta}}(S_i(\hat{\theta})S_i(\hat{\theta})^\intercal)$ is mostly a straightforward symmetric block matrix:

$$
\begin{pmatrix}
\left(\frac{Z_i}{\hat{\xi}} - \frac{1-Z_i}{1-\hat{\xi}}\right)^2 & \begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \left(\frac{Z_i}{\hat{\xi}} - \frac{1-Z_i}{1-\hat{\xi}}\right)\left(\hat{n}_{i,aa} - \pi(X_i;\hat{\zeta}_{aa})\right) & \cdots & \begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \left(\frac{Z_i}{\hat{\xi}} - \frac{1-Z_i}{1-\hat{\xi}}\right)\hat{n}_{i,cc0}\left(Y_i - \eta(X_i;\hat{\beta}_{cc0})\right) & \cdots \\
\hline
 & \begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \left(\hat{n}_{i,aa}\left(1 - 2\pi(X_i;\hat{\zeta}_{aa})\right) + \pi(X_i;\hat{\zeta}_{aa})^2\right) & & \\
 & - \begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal \left(\hat{n}_{i,aa}\ \pi(X_i;\hat{\zeta}_{na}) + \hat{n}_{i,na}\ \pi(X_i;\hat{\zeta}_{aa})\right) & \ddots & E_{\hat{\theta}}\left(S_i(\hat{\zeta})S_i(\hat{\beta})^\intercal\right) \\
 & \vdots & & \\
\hline
 & & \begin{pmatrix}1\\X_i\end{pmatrix}\begin{pmatrix}1\\X_i\end{pmatrix}^\intercal n_{cc0}\left(Y_i - \eta(X_i;\hat{\beta}_{cc0})\right)^2 & \\
 & & 0 & \ddots \\
 & & \vdots &
\end{pmatrix}
$$

where $\hat{n}_{i,stz} = \sum_{u,v} \mathbb{1}_{i,zuv} E_{\hat{\theta}}(C_i = st \mid Z_i = z, D_i^{\mathrm{P}} = u, D_i^{\mathrm{S}} = v, Y_i, X_i)$ and $\hat{n}_{i,st} = \hat{n}_{i,st0} + \hat{n}_{i,st1}$.

Consider the submatrix $E_{\hat{\theta}}\big(S_i(\hat{\zeta})S_i(\hat{\beta})^\intercal\big)$, whose elements are not quite as straightforward. First, consider parameters that correspond to *different* compliance types, e.g. the product of partial derivatives with respect ζ_{an} and β_{ca0}:

$$\begin{aligned}
&\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial \zeta_{an}}\left(\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial \beta_{ca0}}\right) \\
&\qquad = \begin{pmatrix} 1 \\ X_i \end{pmatrix}\begin{pmatrix} 1 \\ X_i \end{pmatrix}^\intercal \big(\mathbb{1}_{i,an010} + \mathbb{1}_{i,an110} - \pi(X_i;\hat{\zeta}_{an})\big)\mathbb{1}_{i,ca001}\big(Y_i - \eta(X_i;\hat{\beta}_{ca0})\big) \\
&\qquad = \begin{pmatrix} 1 \\ X_i \end{pmatrix}\begin{pmatrix} 1 \\ X_i \end{pmatrix}^\intercal \Big(- \pi(X_i;\hat{\zeta}_{an})\mathbb{1}_{i,ca001}\big(Y_i - \eta(X_i;\hat{\beta}_{ca0})\big)\Big).
\end{aligned}$$

Second, consider parameters that correspond to the *same* compliance type, e.g. the product of partial derivatives with respect ζ_{an} and β_{an}:

$$\begin{aligned}
&\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial \zeta_{an}}\left(\frac{\partial(\ell \mid Z_i, D_i^{\mathrm{P}}, D_i^{\mathrm{S}}, Y_i, X_i)}{\partial \beta_{an}}\right) \\
&= \begin{pmatrix} 1 \\ X_i \end{pmatrix}\begin{pmatrix} 1 \\ X_i \end{pmatrix}^\intercal \big(\mathbb{1}_{i,an010} + \mathbb{1}_{i,an110} - \pi(X_i;\hat{\zeta}_{an})\big)(\mathbb{1}_{i,an110} + \mathbb{1}_{i,an010})\big(Y_i - \eta(X_i;\hat{\beta}_{an})\big) \\
&= \begin{pmatrix} 1 \\ X_i \end{pmatrix}\begin{pmatrix} 1 \\ X_i \end{pmatrix}^\intercal \big(\mathbb{1}_{i,an010} + \mathbb{1}_{i,an110}\big)\big(1 - \pi(X_i;\hat{\zeta}_{aa})\big(Y_i - \eta(X_i;\hat{\beta}_{an})\big).
\end{aligned}$$

It then follows that $E_{\hat{\theta}}\big(S_i(\hat{\zeta})S_i(\hat{\beta})^\intercal\big)$ is defined as follows:

$$E_{\hat{\theta}}\big(S_i(\hat{\zeta})S_i(\hat{\beta})^\intercal\big) = \begin{pmatrix} \begin{pmatrix} 1 \\ X_i \end{pmatrix} & \mathbf{0} & \cdots \\ \mathbf{0} & \begin{pmatrix} 1 \\ X_i \end{pmatrix} & \\ \mathbf{0} & \mathbf{0} & \ddots \end{pmatrix} \left[\mathsf{M} - \begin{pmatrix} \pi(X_i;\hat{\zeta}_{aa}) \\ \pi(X_i;\hat{\zeta}_{na}) \\ \pi(X_i;\hat{\zeta}_{an}) \\ \pi(X_i;\hat{\zeta}_{nn}) \\ \pi(X_i;\hat{\zeta}_{cn}) \\ \pi(X_i;\hat{\zeta}_{ca}) \end{pmatrix} \begin{pmatrix} \hat{n}_{i,cc0}\big(Y_i - \eta(X_i;\hat{\beta}_{cc0})\big) \\ \hat{n}_{i,cc1}\big(Y_i - \eta(X_i;\hat{\beta}_{cc1})\big) \\ \hat{n}_{i,cn1}\big(Y_i - \eta(X_i;\hat{\beta}_{cn1})\big) \\ \hat{n}_{i,ca0}\big(Y_i - \eta(X_i;\hat{\beta}_{ca0})\big) \\ (\hat{n}_{i,nn} + \hat{n}_{i,cn0})\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet n})\big) \\ (\hat{n}_{i,aa} + \hat{n}_{i,ca1})\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet a})\big) \\ \hat{n}_{i,an}\big(Y_i - \eta(X_i;\hat{\beta}_{an})\big) \\ \hat{n}_{i,na}\big(Y_i - \eta(X_i;\hat{\beta}_{na})\big) \end{pmatrix}^\intercal \right] \begin{pmatrix} \begin{pmatrix} 1 \\ X_i \end{pmatrix} & \mathbf{0} & \cdots \\ \mathbf{0} & \begin{pmatrix} 1 \\ X_i \end{pmatrix} & \\ \mathbf{0} & \mathbf{0} & \ddots \end{pmatrix}^\intercal$$

where

$$\mathsf{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \hat{n}_{i,aa}\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet a})\big) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{n}_{i,na}\big(Y_i - \eta(X_i;\hat{\beta}_{na})\big) \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{n}_{i,an}\big(Y_i - \eta(X_i;\hat{\beta}_{an})\big) & 0 \\ 0 & 0 & 0 & 0 & \hat{n}_{i,nn}\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet n})\big) & 0 & 0 & 0 \\ 0 & 0 & \hat{n}_{i,cn1}\big(Y_i - \eta(X_i;\hat{\beta}_{cn1})\big) & 0 & \hat{n}_{i,cn1}\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet n})\big) & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{n}_{i,ca0}\big(Y_i - \eta(X_i;\hat{\beta}_{ca0})\big) & 0 & \hat{n}_{i,ca1}\big(Y_i - \eta(X_i;\hat{\beta}_{\bullet a})\big) & 0 & 0 \end{pmatrix}$$

Now that we have the formula for the observed data information matrix, $I_{\mathbf{n}}$, by the delta method an asymptotically consistent estimate of the variance of $ACE(cc|x)_{\text{complete}}$ for a fixed value of x is

$$\widehat{Var}(ACE(cc|x)_{\text{complete}})$$

$$= \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\intercal} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \left.\frac{\partial \eta(x;\beta_{cc0})}{\partial \beta_{cc0}}\right|_{\theta=\hat{\theta}} & \mathbf{0} \\ \mathbf{0} & \left.\frac{\partial \eta(x;\beta_{cc1})}{\partial \beta_{cc1}}\right|_{\theta=\hat{\theta}} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^{\intercal} I_{\mathbf{n}}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \left.\frac{\partial \eta(x;\beta_{cc0})}{\partial \beta_{cc0}}\right|_{\theta=\hat{\theta}} & \mathbf{0} \\ \mathbf{0} & \left.\frac{\partial \eta(x;\beta_{cc1})}{\partial \beta_{cc1}}\right|_{\theta=\hat{\theta}} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\intercal} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \frac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \frac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^{\intercal} I_{\mathbf{n}}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \frac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \frac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Appendix C

APPENDIX FOR CHAPTER 3

C.1 The MCEM Algorithm

The parameter of the model is (θ, Σ), where $\theta = (\xi, \zeta, \beta)$, $\zeta = (\zeta_{nn}, \zeta_{cn})$ and $\beta = (\beta_{cc0}, \beta_{cc1}, \beta_{cn1}, \beta_{\bullet n})$. For subject i of provider j, the observed variables are $(Z_{ji}, D^{\mathrm{P}}_{ji}, D^{\mathrm{S}}_{ji}, X_{ji}, Y_{ji})$ and unobserved variables $(C_{ji}, \epsilon_j, \delta_j)$ where ϵ_j and δ_j are vectors that corresponds to therapist-specific random effects with $\epsilon_j = (\epsilon_{j,cc0}, \epsilon_{j,cc1}, \epsilon_{j,cn1}, \epsilon_{j,\bullet n})^\intercal$ and $\delta_j = (\delta_{j,nn}, \delta_{j,cn})^\intercal$. It is assumed that $(\epsilon_j, \delta_j)^\intercal \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$.

The complete data log-likelihood is:

$$
\begin{aligned}
\ell = \sum_{j=1}^{m} \Bigg[\sum_{i=1}^{n_j} \Bigg\{ & Z_{ji} \log(\xi) + (1 - Z_{ji}) \log(1 - \xi) + \\
& \mathbb{1}_{ji,nn000} \big[Y_{ji} \log \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n})) + \log \pi(X_{ji}; \zeta_{nn}, \delta_{j,nn}) \big] + \\
& \mathbb{1}_{ji,nn100} \big[Y_{ji} \log \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n})) + \log \pi(X_{ji}; \zeta_{nn}, \delta_{j,nn}) \big] + \\
& \mathbb{1}_{ji,cn000} \big[Y_{ji} \log \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n})) + \log \pi(X_{ji}; \zeta_{cn}, \delta_{j,cn}) \big] + \\
& \mathbb{1}_{ji,cn110} \big[Y_{ji} \log \eta(X_{ji}; \beta_{cn1}, \eta_{j,cn1}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{cn1}, \eta_{j,cn1})) + \log \pi(X_{ji}; \zeta_{cn}, \delta_{j,cn}) \big] + \\
& \mathbb{1}_{ji,cc000} \big[Y_{ji} \log \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0})) + \log \pi(X_{ji}; \zeta_{cc}, \delta_{j,cc}) \big] + \\
& \mathbb{1}_{ji,cc111} \big[Y_{ji} \log \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1}) + (1 - Y_{ji}) \log(1 - \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1})) + \log \pi(X_{ji}; \zeta_{cc}, \delta_{j,cc}) \big] \Bigg\} + \\
& \log \phi \left(\begin{pmatrix} \epsilon_j^q \\ \delta_j^q \end{pmatrix} \middle| \Sigma \right) \Bigg],
\end{aligned}
$$

where $\mathbb{1}_{ji,stzuv} = \mathbb{1}(C_{ji} = st, Z_{ji} = z, D^{\mathrm{P}}_{ji} = u, D^{\mathrm{S}}_{ji} = v)$.

C.1.1 E-step

For the E-step of the MCEM-algorithm, choose initial values for (θ, Σ), and denote them by $(\theta^{(0)}, \Sigma^{(0)})$. In this step we consider the marginal expectation of the complete data log-likelihood conditional on the observed data with respect to the conditional distribution of $(\epsilon_1, ..., \epsilon_m, \delta_1, ..., \delta_m) \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}$ under the current parameter value, $(\theta^{(0)}, \Sigma^{(0)})$:

$$
\begin{aligned}
&E_{\theta^{(0)},\Sigma^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) = \sum_{j=1}^{m} \int \left[\sum_{i=1}^{n_j} \Big\{ \big[\mathbb{1}(Z_{ji} = 1) \log(\xi) + \mathbb{1}(Z_{ji} = 0) \log(1 - \xi) \big] + \right. \\
&(n^{(0)}_{ji,cn000} + n^{(0)}_{ji,cn110}) \log \pi(X_{ji}; \zeta_{cn}, \delta_{j,cn}) + (n^{(0)}_{ji,nn000} + n^{(0)}_{ji,nn100}) \log \pi(X_{ji}; \zeta_{nn}, \delta_{j,nn}) + \\
&n^{(0)}_{ji,cc0001} \log \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0}) + n^{(0)}_{ji,cc0000} \log(1 - \eta(X_{ji}; \beta_{cc0}, \epsilon_{j,cc0})) + \\
&n^{(0)}_{ji,cc1111} \log \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1}) + n^{(0)}_{ji,cc1110} \log(1 - \eta(X_{ji}; \beta_{cc1}, \epsilon_{j,cc1})) + \\
&n^{(0)}_{ji,cn1101} \log \eta(X_{ji}; \beta_{cn1}, \epsilon_{j,cn1}) + n^{(0)}_{ji,cn1100} \log(1 - \eta(X_{ji}; \beta_{cn1}, \epsilon_{j,cn1})) + \\
&(n^{(0)}_{ji,nn0001} + n^{(0)}_{ji,nn1001} + n^{(0)}_{ji,cn0001}) \log \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n}) + \\
&\left. (n^{(0)}_{ji,nn0000} + n^{(0)}_{ji,nn1000} + n^{(0)}_{ji,cn0000}) \log(1 - \eta(X_{ji}; \beta_{\bullet n}, \epsilon_{j,\bullet n})) \Big\} + \right. \\
&\left. \log \phi \left(\begin{pmatrix} \epsilon_j^q \\ \delta_j^q \end{pmatrix} \middle| \Sigma \right) \right] \, f\big((\epsilon_j, \delta_j)^{\intercal} \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}, \theta^{(0)}, \Sigma^{(0)}\big) \, d\big((\epsilon_j, \delta_j)^{\intercal}\big),
\end{aligned}
$$

where $n^{(0)}_{ji,stzuvy} \equiv P_{\theta^{(0)}}(C_{ji} = st, Z_{ji} = z, D^{\mathrm{P}}_{ji} = u, D^{\mathrm{S}}_{ji} = v, Y_{ji} = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ and $n^{(0)}_{ji,stzuv} = n^{(0)}_{ji,stzuv0} + n^{(0)}_{ji,stzuv1}$. Consider $n^{(0)}_{ji,st000y}$ for $st \in \{nn, cn, cc\}$.

$$
\begin{aligned}
n^{(0)}_{ji,st000y} &= P_{\theta^{(0)}}(C_{ji} = st, Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}) \\
&= P_{\theta^{(0)}}(C_{ji} = st, Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y \mid Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y, X_{ji}) \\
&= \mathbb{1}(Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y) P_{\theta^{(0)}}(C_{ji} = st \mid Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y, X_{ji}).
\end{aligned}
$$

By Baye's rule,

$$
\begin{aligned}
&P_{\theta^{(0)}}(C_{ji} = st \mid Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0, Y_{ji} = y, X_{ji}) \\
&\quad = \frac{P_{\theta^{(0)}}(Y_{ji} = y, Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0 \mid C_{ji} = st, X_{ji}) P_{\theta^{(0)}}(C_{ji} = st \mid X_{ji})}{P_{\theta^{(0)}}(Y_{ji} = y, Z_{ji} = 0, D^{\mathrm{P}}_{ji} = 0, D^{\mathrm{S}}_{ji} = 0 \mid X_{ji})} \\
&\quad = \frac{P_{\theta^{(0)}}(Y_{ji} = y \mid Z_{ji} = 0, C_{ji} = st, X_{ji}) P_{\theta^{(0)}}(C_{ji} = st \mid X_{ji})}{\sum_{st=\{nn,cn,cc\}} P_{\theta^{(0)}}(Y_{ji} = y \mid Z_{ji} = 0, C_{ji} = st, X_{ji}) P_{\theta^{(0)}}(C_{ji} = st \mid X_{ji})}
\end{aligned}
$$

where

$$
\begin{aligned}
P_{\theta^{(0)}}(Y_{ji} = y \mid Z_{ji} = 0, C_{ji} = nn, X_{ji}) &= \eta(X_{ji}; \beta^{(0)}_{\bullet n}, \epsilon_{j,nn})^y \big(1 - \eta(X_{ji}; \beta^{(0)}_{\bullet n}, \epsilon_{j,\bullet n})\big)^{1-y}, \\
P_{\theta^{(0)}}(Y_{ji} = y \mid Z_{ji} = 0, C_{ji} = cn, X_{ji}) &= \eta(X_{ji}; \beta^{(0)}_{\bullet n}, \epsilon_{j,\bullet n})^y \big(1 - \eta(X_{ji}; \beta^{(0)}_{\bullet n}, \epsilon_{j,\bullet n})\big)^{1-y}, \\
P_{\theta^{(0)}}(Y_{ji} = y \mid Z_{ji} = 0, C_{ji} = cc, X_{ji}) &= \eta(X_{ji}; \beta^{(0)}_{cc0}, \epsilon_{j,cc0})^y \big(1 - \eta(X_{ji}; \beta^{(0)}_{cc0}, \epsilon_{j,cc0})\big)^{1-y}, \\
P_{\theta^{(0)}}(C_{ji} = nn \mid X_{ji}) &= \pi(X_{ji}; \zeta^{(0)}_{nn}, \delta_{j,nn}), \\
P_{\theta^{(0)}}(C_{ji} = cn \mid X_{ji}) &= \pi(X_{ji}; \zeta^{(0)}_{cn}, \delta_{j,cn}), \\
P_{\theta^{(0)}}(C_{ji} = cc \mid X_{ji}) &= \pi(X_{ji}; \zeta^{(0)}_{cc}, \delta_{j,cc}).
\end{aligned}
$$

Lastly,

$$
n^{(0)}_{ji,cc111y} = n^{(0)}_{ji,cn110y} = n^{(0)}_{ji,nn100y} = \mathbb{1}(Y_{ji} = y).
$$

C.1.2 Monte Carlo Approximation

To approximate the integrals in $E_{\theta^{(0)},\Sigma^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$, we use Monte Carlo approximation [Wei and Tanner, 1990]. Specifically, we implement the automated Monte Carlo EM algorithm for estimation in GLMMs proposed by Booth and Hobert [1999].

The expectation for the E-step is approximated by:

$$E^{MC}_{\theta^{(0)},\Sigma^{(0)}}(\ell \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{S}},\mathbf{Y},\mathbf{X}) = \frac{1}{Q}\sum_{j=1}^{m}\sum_{q=1}^{Q}\Bigg[\sum_{i=1}^{n_j}\Big\{\mathbb{1}(Z_{ji}=1)\log(\xi)+\mathbb{1}(Z_{ji}=0)\log(1-\xi)+$$
$$(n^{(0,q)}_{ji,nn0}+n^{(0,q)}_{ji,nn1})\log\pi(X_{ji};\zeta_{nn},\delta^q_{j,nn})+(n^{(0,q)}_{ji,cn0}+n^{(0,q)}_{ji,cn1})\log\pi(X_{ji};\zeta_{cn},\delta^q_{j,cn})+$$
$$n^{(0,q)}_{ji,cc0001}\log\eta(X_{ji};\beta_{cc0},\epsilon^q_{j,cc0})+n^{(0,q)}_{ji,cc0000}\log(1-\eta(X_{ji};\beta_{cc0},\epsilon^q_{j,cc0}))+$$
$$n^{(0,q)}_{ji,cc1111}\log\eta(X_{ji};\beta_{cc1},\epsilon^q_{j,cc1})+n^{(0,q)}_{ji,cc1110}\log(1-\eta(X_{ji};\beta_{cc1},\epsilon^q_{j,cc1}))+$$
$$n^{(0,q)}_{ji,cn1101}\log\eta(X_{ji};\beta_{cn1},\epsilon^q_{j,cn1})+n^{(0,q)}_{ji,cn1100}\log(1-\eta(X_{ji};\beta_{cn1},\epsilon^q_{j,cn1}))+$$
$$(n^{(0,q)}_{ji,nn0001}+n^{(0,q)}_{ji,nn1001}+n^{(0,q)}_{ji,cn0001})\log\eta(X_{ji};\beta_{\bullet n},\epsilon^q_{j,\bullet n})+$$
$$(n^{(0,q)}_{ji,nn0000}+n^{(0,q)}_{ji,nn1000}+n^{(0,q)}_{ji,cn0000})\log(1-\eta(X_{ji};\beta_{\bullet n},\epsilon^q_{j,\bullet n}))\Big\}+$$
$$\log\phi\left(\begin{pmatrix}\epsilon^q_j\\ \delta^q_j\end{pmatrix}\middle|\,\Sigma\right)\Bigg],$$

where ϵ^q_j,δ^q_j are random draws from $f\big((\epsilon_j,\delta_j)^\intercal \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{S}},\mathbf{Y},\mathbf{X},\theta^{(0)},\Sigma^{(0)}\big)$ and $n^{(0,q)}_{ji,stzuvy}$ is defined as follows.

$$n^{(0,q)}_{ji,st000y} = \frac{\mathbb{1}(Z_{ji}=0,D^{\mathrm{P}}_{ji}=0,D^{\mathrm{S}}_{ji}=0,Y_{ji}=y)P_{\theta^{(0,q)}}(Y_{ji}=y\mid Z_{ji}=0,C_{ji}=st,X_{ji}=x)P_{\theta^{(0,q)}}(C_{ji}=nn\mid X_{ji}=x)}{\sum_{st=\{nn,cn,cc\}}P_{\theta^{(0,q)}}(Y_{ji}=y\mid Z_{ji}=0,C_{ji}=st,X_{ji}=x)P_{\theta^{(0,q)}}(C_{ji}=st\mid X_{ji}=x)}$$

for $y\in\{0,1\}$ where

$$\begin{aligned}
P_{\theta^{(0,q)}}(Y_{ji}=y\mid Z_{ji}=0,C_{ji}=nn,X_{ji}) &= \eta(X_{ji};\beta^{(0)}_{\bullet n},\epsilon^q_{j,nn})^y\big(1-\eta(X_{ji};\beta^{(0)}_{\bullet n},\epsilon^q_{j,\bullet n})\big)^{1-y},\\
P_{\theta^{(0,q)}}(Y_{ji}=y\mid Z_{ji}=0,C_{ji}=cn,X_{ji}) &= \eta(X_{ji};\beta^{(0)}_{\bullet n},\epsilon^q_{j,\bullet n})^y\big(1-\eta(X_{ji};\beta^{(0)}_{\bullet n},\epsilon^q_{j,\bullet n})\big)^{1-y},\\
P_{\theta^{(0,q)}}(Y_{ji}=y\mid Z_{ji}=0,C_{ji}=cc,X_{ji}) &= \eta(X_{ji};\beta^{(0)}_{cc0},\epsilon^q_{j,cc0})^y\big(1-\eta(X_{ji};\beta^{(0)}_{cc0},\epsilon^q_{j,cc0})\big)^{1-y},\\
P_{\theta^{(0,q)}}(C_{ji}=nn\mid X_{ji}) &= \pi(X_{ji};\zeta^{(0)}_{nn},\delta^q_{j,nn}),\\
P_{\theta^{(0,q)}}(C_{ji}=cn\mid X_{ji}) &= \pi(X_{ji};\zeta^{(0)}_{cn},\delta^q_{j,cn}),\\
P_{\theta^{(0,q)}}(C_{ji}=cc\mid X_{ji}) &= \pi(X_{ji};\zeta^{(0)}_{cc},\delta^q_{j,cc})
\end{aligned}$$

and $n^{(0,q)}_{ji,cc111y}=n^{(0,q)}_{ji,cn110y}=n^{(0,q)}_{ji,nn100y}=\mathbb{1}(Y_{ji}=y)$.

For each j, the random draws from $f\big((\epsilon_j,\delta_j)^\intercal \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{S}},\mathbf{Y},\mathbf{X},\theta^{(0)},\Sigma^{(0)}\big)$ are obtained via rejection sampling [Geweke, 1996]. Note that

$$f\big((\epsilon_j,\delta_j)^\intercal \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{S}},\mathbf{Y},\mathbf{X},\theta^{(0)},\Sigma^{(0)}\big)\propto f(\mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{S}},\mathbf{Y},\mathbf{X}\mid\theta^{(0)},\epsilon_j,\delta_j)\ \phi\big((\epsilon_j,\delta_j)^\intercal\big|\Sigma^{(0)}\big),$$

where

$$
\begin{aligned}
&f(\mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X} \mid \theta^{(0)}, \epsilon_j, \delta_j) = \\
&\prod_{i=1}^{n_j} \Big[(1-\xi^{(0)})\ \pi(X_{ji}; \zeta_{nn}^{(0)}, \delta_{j,nn})\Big]^{\mathbb{1}_{ji,nn000}}\ \big(1-\eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,nn0000}}\ \eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})^{\mathbb{1}_{ji,nn0001}}\ \times \\
&\Big[\xi^{(0)}\ \pi(X_{ji}; \zeta_{nn}^{(0)}, \delta_{j,nn})\Big]^{\mathbb{1}_{ji,nn100}}\ \big(1-\eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,nn1000}}\ \eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})^{\mathbb{1}_{ji,nn1001}}\ \times \\
&\Big[(1-\xi^{(0)})\ \pi(X_{ji}; \zeta_{cn}^{(0)}, \delta_{j,cn})\Big]^{\mathbb{1}_{ji,cn000}}\ \big(1-\eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})\big)^{\mathbb{1}_{ji,cn0000}}\ \eta(X_{ji}; \beta_{\bullet n}^{(0)}, \epsilon_{j,\bullet n})^{\mathbb{1}_{ji,cn0001}}\ \times \\
&\Big[\xi^{(0)}\ \pi(X_{ji}; \zeta_{cn}^{(0)}, \delta_{j,cn})\Big]^{\mathbb{1}_{ji,cn110}}\ \big(1-\eta(X_{ji}; \beta_{cn1}^{(0)}, \epsilon_{j,cn1})\big)^{\mathbb{1}_{ji,cn1100}}\ \eta(X_{ji}; \beta_{cn1}, \epsilon_{j,cn1}^{(0)})^{\mathbb{1}_{ji,cn1101}}\ \times \\
&\Big[(1-\xi^{(0)})\ \pi(X_{ji}; \zeta_{cc}^{(0)}, \delta_{j,cc})\Big]^{\mathbb{1}_{ji,cc000}}\ \big(1-\eta(X_{ji}; \beta_{cc0}^{(0)}, \epsilon_{j,cc0})\big)^{\mathbb{1}_{ji,cc0000}}\ \eta(X_{ji}; \beta_{cc0}^{(0)}, \epsilon_{j,cc0})^{\mathbb{1}_{ji,cc0001}}\ \times \\
&\Big[\xi\ \pi(X_{ji}; \zeta_{cc}^{(0)}, \delta_{j,cc})\Big]^{\mathbb{1}_{ji,cc111}}\ \big(1-\eta(X_{ji}; \beta_{cc1}^{(0)}, \epsilon_{j,cc1})\big)^{\mathbb{1}_{ji,cc1110}}\ \eta(X_{ji}; \beta_{cc1}^{(0)}, \epsilon_{j,cc1})^{\mathbb{1}_{ji,cc1111}}.
\end{aligned}
$$

The rejection sampling procedure is desribed as follows:

- Step 1: sample $(\epsilon_j, \delta_j)^{\intercal}$ from $\mathcal{N}(\mathbf{0}, \Sigma^{(0)})$ and independently sample w from the uniform(0,1) distribution.

- Step 2: if $w \leq \frac{f(\mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X} | \theta^{(0)}, \epsilon_j, \delta_j)}{\tau}$ where $\tau = \sup_{\epsilon_j, \delta_j} f(\mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X} \mid \theta^{(0)}, \epsilon_j, \delta_j)$, then accept $(\epsilon_j, \delta_j)^{\intercal}$; if not, go to step 1. τ is esimated numerically, in practice this is achieved by using standard optimization functions in R (e.g. the `nlm` function, which utilizes a Newton-type algorithm).

For brevity, from this point forward we let $n_{ji,nn1}^{(0,q)} \equiv \sum_{u,v \in \{0,1\}} n_{ji,nn1uv}^{(0,q)} = n_{ji,nn100}^{(0,q)}$ and $n_{ji,nn0}^{(0,q)} \equiv \sum_{u,v \in \{0,1\}} n_{ji,nn0uv}^{(0,q)} = n_{ji,nn000}^{(0,q)}$. For $st \in \{cn, cc\}$ and $z \in \{0,1\}$ we define $n_{ji,stz}^{(0)}$ similarly.

C.1.3 M-step

The M-step consists of maximizing $E_{\theta^{(0)}, \Sigma^{(0)}}^{MC}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ with respect to (θ, Σ). Since θ and Σ occur separately in two terms in $E_{\theta^{(0)}, \Sigma^{(0)}}^{MC}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$, each term can be maximized separately.

Update for Σ

The term to be maximized in $E^{MC}_{\theta^{(0)},\Sigma^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ with respect to Σ is simply:

$$\frac{1}{Q}\sum_{j=1}^{m}\sum_{q=1}^{Q}\log\phi\left(\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}\middle|\Sigma\right).$$

Since the above summation is the log likelihood of the joint distribution of mQ iid draws from a multivariate normal distribution, the MLE is well known from multivariate statistical theory and the update for Σ is:

$$\Sigma^{(k+1)} = \frac{1}{mQ}\sum_{j=1}^{m}\sum_{q=1}^{Q}\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}^{\intercal}.$$

Update for θ

No closed-form solution exists for the term to be maximized in $E^{MC}_{\theta^{(0)},\Sigma^{(0)}}(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X})$ with respect to θ:

$$\begin{aligned}
&\frac{1}{Q}\sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_j}\Big\{\mathbb{1}(Z_{ji}=1)\log(\xi)+\mathbb{1}(Z_{ji}=0)\log(1-\xi)+\\
&(n^{(0,q)}_{ji,nn0}+n^{(0,q)}_{ji,nn1})\log\pi(X_{ji};\zeta_{nn},\delta^q_{j,nn})+(n^{(0,q)}_{ji,cn0}+n^{(0,q)}_{ji,cn1})\log\pi(X_{ji};\zeta_{cn},\delta^q_{j,cn})+\\
&n^{(0,q)}_{ji,cc0001}\log\eta(X_{ji};\beta_{cc0},\epsilon^q_{j,cc0})+n^{(0,q)}_{ji,cc0000}\log(1-\eta(X_{ji};\beta_{cc0},\epsilon^q_{j,cc0}))+\\
&n^{(0,q)}_{ji,cc1111}\log\eta(X_{ji};\beta_{cc1},\epsilon^q_{j,cc1})+n^{(0,q)}_{ji,cc1110}\log(1-\eta(X_{ji};\beta_{cc1},\epsilon^q_{j,cc1}))+\\
&n^{(0,q)}_{ji,cn1101}\log\eta(X_{ji};\beta_{cn1},\epsilon^q_{j,cn1})+n^{(0,q)}_{ji,cn1100}\log(1-\eta(X_{ji};\beta_{cn1},\epsilon^q_{j,cn1}))+\\
&(n^{(0,q)}_{ji,nn0001}+n^{(0,q)}_{ji,nn1001}+n^{(0,q)}_{ji,cn0001})\log\eta(X_{ji};\beta_{\bullet n},\epsilon^q_{j,\bullet n})+\\
&(n^{(0,q)}_{ji,nn0000}+n^{(0,q)}_{ji,nn1000}+n^{(0,q)}_{ji,cn0000})\log(1-\eta(X_{ji};\beta_{\bullet n},\epsilon^q_{j,\bullet n}))\Big\}.
\end{aligned}\tag{C.1}$$

However, (C.1) is similar to the expectation in the E-step of the GEM-algorithm in Chapter 2 and hence the update for θ is obtained similar to the M-step in Chapter 2. The update $\theta^{(k+1)}$ is given by the following:

$$\theta^{(k+1)} = \theta^{(k)} + \left(-\frac{\partial^2}{\partial\theta\partial\theta^{\intercal}}\Big(\text{expression (C.1)}\Big)\Big|_{\theta=\theta^{(k)}}\right)^{-1}\left(\frac{\partial}{\partial\theta}\Big(\text{expression (C.1)}\Big)\Big|_{\theta=\theta^{(k)}}\right).$$

The score of (C.1), $\frac{\partial}{\partial\theta}\big(\text{expression (C.1)}\big)|_{\theta=\theta^{(k)}}$, is:

$$\sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_j}\begin{pmatrix} \begin{pmatrix} \frac{Z_{ji}}{\xi^{(k)}} - \frac{1-Z_{ji}}{1-\xi^{(k)}} \end{pmatrix} \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix}\big(n^{(k,q)}_{ji,nn0} + n^{(k,q)}_{ji,nn1} - \pi(X_{ji};\zeta^{(k)}_{nn},\delta^{q}_{j,nn})\big) \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix}\big(n^{(k,q)}_{ji,cn0} + n^{(k,q)}_{ji,cn1} - \pi(X_{ji};\zeta^{(k)}_{cn},\delta^{q}_{j,cn})\big) \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} n^{(k,q)}_{ji,cc0}\big(Y_{ji} - \eta(X_{ji};\beta^{(k)}_{cc0},\epsilon^{q}_{j,cc0})\big) \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} n^{(k,q)}_{ji,cc1}\big(Y_{ji} - \eta(X_{ji};\beta^{(k)}_{cc1},\epsilon^{q}_{j,cc1})\big) \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} n^{(k,q)}_{ji,cn1}\big(Y_{ji} - \eta(X_{ji};\beta^{(k)}_{cn1},\epsilon^{q}_{j,cn1})\big) \\ \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix}\big(n^{(k,q)}_{ji,nn0} + n^{(k,q)}_{ji,cn0} + n^{(k,q)}_{ji,cc0}\big)\big(Y_{ji} - \eta(X_{ji};\beta^{(k)}_{\bullet n},\epsilon^{q}_{j,\bullet n})\big) \end{pmatrix}.$$

The negative Hessian of (C.1), $-\frac{\partial^2}{\partial\theta\partial\theta^\intercal}\big(\text{expression (C.1)}\big)|_{\theta=\theta^{(k)}}$, is given by the following block diagonal matrix:

$$\begin{pmatrix} \frac{\partial^2}{\partial\xi^2}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} & \mathbf{0}^\intercal & \mathbf{0}^\intercal \\ \mathbf{0} & \frac{\partial^2}{\partial\zeta\partial\zeta^\intercal}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} & \mathbf{0}\mathbf{0}^\intercal \\ \mathbf{0} & \mathbf{0}\mathbf{0}^\intercal & \frac{\partial^2}{\partial\beta\partial\beta^\intercal}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} \end{pmatrix}$$

where

$$\frac{\partial^2}{\partial\xi^2}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} = \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\left(\frac{Z_{ji}}{(\xi^{(k)})^2}+\frac{1-Z_{ji}}{(1-\xi^{(k)})^2}\right),$$

$$\frac{\partial^2}{\partial\zeta\partial\zeta^\intercal}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} = \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\begin{pmatrix} \begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal\Big(\pi(X_{ji};\zeta_{nn}^{(k)},\delta_{j,nn}^{q})-\pi(X_{ji};\zeta_{nn}^{(k)},\delta_{j,nn}^{q})^2\Big) & \\ -\begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal\pi(X_{ji};\zeta_{cn}^{(k)},\delta_{j,cn}^{q})\pi(X_{ji};\zeta_{cn}^{(k)},\delta_{j,cn}^{q}) & \ddots \end{pmatrix},$$

$$\frac{\partial^2}{\partial\beta\partial\beta^\intercal}\big(\text{expression (C.1)}\big)\Big|_{\theta=\theta^{(k)}} = \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\begin{pmatrix} \begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal n_{ji,cc0}^{(k,q)}\Big(\eta(X_{ji};\beta_{cc0}^{(k)},\epsilon_{j,cc0}^{q})-\eta(X_{ji};\beta_{cc0}^{(k)},\epsilon_{j,cc0}^{q})^2\Big) & \\ 0 & \ddots \\ \vdots & \end{pmatrix}.$$

The E- and M-steps are repeated until the updates converge. Since the EM algorithm only guarantees a local maxima is reached, we repeat the algorithm at multiple starting values. We try many different initial values $(\theta^{(0)}, \Sigma^{(0)})$ and choose the solution that corresponds to the highest converged likelihood value, which we denote by $(\hat{\theta}, \hat{\Sigma})$. The ML estimate of $ACE(cc|x)$ is

$$ACE(cc|x)_{\text{cluster}} = \eta(x; \hat{\beta}_{cc1}) - \eta(x; \hat{\beta}_{cc0}).$$

C.1.4 Variance

The asymptotic variance of $ACE(cc|x)_{\text{complete}}$ is obtained from the inverse observed data information matrix, $I_{\mathbf{n}}$. We again employ Louis's formula [Louis, 1982] to obtain $I_{\mathbf{n}}$. Note that we are only concerned with the unique elements of Σ as parameters, and so we denote the half-vectorization of Σ by $\vec{\Sigma}$.

$$\begin{aligned}
I_{\mathbf{n}} &= E_{\hat{\theta},\hat{\Sigma}}\Big(-\frac{\partial^2(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})\partial(\theta, \vec{\Sigma})^{\intercal}}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) - Cov_{\hat{\theta},\hat{\Sigma}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) \\
&= E_{\hat{\theta},\hat{\Sigma}}\Big(-\frac{\partial^2(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})\partial(\theta, \vec{\Sigma})^{\intercal}}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) + \\
&\quad E_{\hat{\theta},\hat{\Sigma}}\Big(\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big)\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big)^{\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) - \\
&\quad E_{\hat{\theta},\hat{\Sigma}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) E_{\hat{\theta},\hat{\Sigma}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big)^{\intercal}.
\end{aligned}$$

Since the expectations are taken with respect to the conditional distribution $f\big((\epsilon_j, \delta_j)^{\intercal} \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{S}}, \mathbf{Y}, \mathbf{X}, \hat{\theta}, \hat{\Sigma}\big)$, we approximate $I_{\mathbf{n}}$ with the Monte Carlo sample of random effects from the last iteration of the MCEM algorithm:

$$\begin{aligned}
I_{\mathbf{n}} &\approx E^{MC}_{\hat{\theta},\hat{\Sigma}}\Big(-\frac{\partial^2(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})\partial(\theta, \vec{\Sigma})^{\intercal}}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) + && \text{(C.2)} \\
&\quad E^{MC}_{\hat{\theta},\hat{\Sigma}}\Big(\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big)\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big)^{\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) - && \text{(C.3)} \\
&\quad E^{MC}_{\hat{\theta},\hat{\Sigma}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big) E_{\hat{\theta},\hat{\Sigma}}\Big(\frac{\partial(\ell \mid \mathbf{Z}, \mathbf{D}^{\mathrm{P}}, \mathbf{D}^{\mathrm{P}}, \mathbf{Y}, \mathbf{X})}{\partial(\theta, \vec{\Sigma})}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\Big)^{\intercal}. && \text{(C.4)}
\end{aligned}$$

The expressions in (C.2) and (C.4) were partially derived in Section C.1.3. First, $E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(\left.\frac{\partial(\ell|\mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial(\theta,\vec{\Sigma})}\right|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right)$ is given by the following:[1]

$$\sum_{j=1}^{m}\sum_{q=1}^{Q}\begin{pmatrix}
\sum_{i=1}^{n_j}\left(\frac{Z_{ji}}{\hat{\xi}}-\frac{1-Z_{ji}}{1-\hat{\xi}}\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,nn0}+\hat{n}^q_{ji,nn1}-\pi(X_{ji};\hat{\zeta}_{nn},\delta^q_{j,nn})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,cn0}+\hat{n}^q_{ji,cn1}-\pi(X_{ji};\hat{\zeta}_{cn},\delta^q_{j,cn})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\hat{n}^q_{ji,cc0}\left(Y_{ji}-\eta(X_{ji};\hat{\beta}_{cc0},\epsilon^q_{j,cc0})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\hat{n}^q_{ji,cc1}\left(Y_{ji}-\eta(X_{ji};\hat{\beta}_{cc1},\epsilon^q_{j,cc1})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\hat{n}^q_{ji,cn1}\left(Y_{ji}-\eta(X_{ji};\hat{\beta}_{cn1},\epsilon^q_{j,cn1})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,nn0}+\hat{n}^q_{ji,cn0}+\hat{n}^q_{ji,cc0}\right)\left(Y_{ji}-\eta(X_{ji};\hat{\beta}_{\bullet n},\epsilon^q_{j,\bullet n})\right)\\[2ex]
\operatorname{vech}\left(-\frac{1}{2}\left(\hat{\Sigma}^{-1}-\hat{\Sigma}^{-1}\begin{pmatrix}\epsilon^q_j\\\delta^q_j\end{pmatrix}\begin{pmatrix}\epsilon^q_j\\\delta^q_j\end{pmatrix}^{\intercal}\hat{\Sigma}^{-1}\right)\right)
\end{pmatrix}.$$

where $\hat{n}^q_{ji,stz}=n^{(k,q)}_{ji,stz}$ where k is the final iteration of the MCEM algorithm.

[1] vech denotes the half-vectorization operator

Next, $E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(- \frac{\partial^2(\ell|\mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial(\theta,\vec{\Sigma})\partial(\theta,\vec{\Sigma})^\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right)$ is given by the block diagonal matrix

$$\begin{pmatrix} \mathbf{A} & & & 0 \\ & \mathbf{B} & & \\ & & \mathbf{C} & \\ 0 & & & \mathbf{D} \end{pmatrix}$$

with elements:

$$\begin{aligned}
\mathbf{A} &= E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(- \frac{\partial^2(\ell \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial\xi^2}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right) \\
&= \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\left(\frac{Z_{ji}}{(\hat{\xi})^2} + \frac{1-Z_{ji}}{(1-\hat{\xi})^2}\right) \\
\mathbf{B} &= E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(- \frac{\partial^2(\ell \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial\zeta\partial\zeta^\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right) \\
&= \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\begin{pmatrix} \begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal \left(\pi(X_{ji};\hat{\zeta}_{nn},\delta^q_{j,nn}) - \pi(X_{ji};\hat{\zeta}_{nn},\delta^q_{j,nn})^2\right) & \\ -\begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal \pi(X_{ji};\hat{\zeta}_{cn},\delta^q_{j,cn})\pi(X_{ji};\hat{\zeta}_{cn},\delta^q_{j,cn}) & \ddots \end{pmatrix} \\
\mathbf{C} &= E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(- \frac{\partial^2(\ell \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial\beta\partial\beta^\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right) \\
&= \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_i}\begin{pmatrix} \begin{pmatrix}1\\X_{ji}\end{pmatrix}\begin{pmatrix}1\\X_{ji}\end{pmatrix}^\intercal \hat{n}^q_{ji,cc0}\left(\eta(X_{ji};\hat{\beta}_{cc0},\epsilon^q_{j,cc0}) - \eta(X_{ji};\hat{\beta}_{cc0},\epsilon^q_{j,cc0})^2\right) & \\ 0 & \ddots \\ \vdots & \end{pmatrix} \\
\mathbf{D} &= E^{MC}_{\hat{\theta},\hat{\Sigma}}\left(- \frac{\partial^2(\ell \mid \mathbf{Z},\mathbf{D}^{\mathrm{P}},\mathbf{D}^{\mathrm{P}},\mathbf{Y},\mathbf{X})}{\partial\vec{\Sigma}\partial\vec{\Sigma}^\intercal}\Big|_{(\theta,\Sigma)=(\hat{\theta},\hat{\Sigma})}\right)
\end{aligned}$$

Now consider (C.3), which is given by the block diagonal matrix

$$\begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F}^{\mathsf{T}} & \mathbf{G} \end{pmatrix}$$

with elements defined as follows.

$$\mathbf{E} = \sum_{j=1}^{m}\sum_{q=1}^{Q}\sum_{i=1}^{n_j} \begin{pmatrix} \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} & \mathbf{0} & \cdots \\ \mathbf{0} & \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} & \\ \mathbf{0} & \mathbf{0} & \ddots \end{pmatrix} \left[\mathsf{M} - \begin{pmatrix} \pi(X_{ji};\hat{\zeta}_{nn},\delta^q_{j,nn}) \\ \pi(X_{ji};\hat{\zeta}_{cn},\delta^q_{j,cn}) \end{pmatrix} \begin{pmatrix} \hat{n}^q_{ji,cc0}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cc0},\epsilon^q_{j,cc0})\big) \\ \hat{n}^q_{ji,cc1}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cc1},\epsilon^q_{j,cc1})\big) \\ \hat{n}^q_{ji,cn1}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cn1},\epsilon^q_{j,cn1})\big) \\ (\hat{n}^q_{ji,nn} + \hat{n}^q_{ji,cn0})\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{\bullet n},\epsilon^q_{j,\bullet n})\big) \end{pmatrix}^\intercal \right] \begin{pmatrix} \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} & \mathbf{0} & \cdots \\ \mathbf{0} & \begin{pmatrix} 1 \\ X_{ji} \end{pmatrix} & \\ \mathbf{0} & \mathbf{0} & \ddots \end{pmatrix}^\intercal$$

where

$$\mathsf{M} = \begin{pmatrix} 0 & 0 & 0 & \hat{n}^q_{ji,nn}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{\bullet n},\epsilon^q_{j,\bullet n})\big) \\ 0 & 0 & \hat{n}^q_{ji,cn1}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cn1},\epsilon^q_{j,cn1})\big) & \hat{n}^q_{ji,cn1}\big(Y_{ji} - \eta(X_{ji};\hat{\beta}_{\bullet n},\epsilon^q_{j,\bullet n})\big) \end{pmatrix}.$$

$$\mathbf{G} = \sum_{j=1}^{m}\sum_{q=1}^{Q} \left[\operatorname{vech}\left(-\tfrac{1}{2}\left(\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} \begin{pmatrix} \epsilon^q_j \\ \delta^q_j \end{pmatrix} \begin{pmatrix} \epsilon^q_j \\ \delta^q_j \end{pmatrix}^\intercal \hat{\Sigma}^{-1} \right) \right) \right] \left[\operatorname{vech}\left(-\tfrac{1}{2}\left(\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} \begin{pmatrix} \epsilon^q_j \\ \delta^q_j \end{pmatrix} \begin{pmatrix} \epsilon^q_j \\ \delta^q_j \end{pmatrix}^\intercal \hat{\Sigma}^{-1} \right) \right) \right]^\intercal.$$

$$\mathbf{F} = \sum_{j=1}^{m}\sum_{q=1}^{Q} \operatorname{vech}\left(-\tfrac{1}{2}\left(\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}\begin{pmatrix}\epsilon_j^q\\ \delta_j^q\end{pmatrix}^{\intercal}\hat{\Sigma}^{-1}\right)\right)$$

$$\begin{pmatrix}
\sum_{i=1}^{n_j}\left(\frac{Z_{ji}}{\hat{\xi}} - \frac{1-Z_{ji}}{1-\hat{\xi}}\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,nn0} + \hat{n}^q_{ji,nn1} - \pi(X_{ji};\hat{\zeta}_{nn},\delta^q_{j,nn})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,cn0} + \hat{n}^q_{ji,cn1} - \pi(X_{ji};\hat{\zeta}_{cn},\delta^q_{j,cn})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\hat{n}^q_{ji,cc0}\left(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cc0},\epsilon^q_{j,cc0})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\hat{n}^q_{ji,cc1}\left(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cc1},\epsilon^q_{j,cc1})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\hat{n}^q_{ji,cn1}\left(Y_{ji} - \eta(X_{ji};\hat{\beta}_{cn1},\epsilon^q_{j,cn1})\right)\\[2ex]
\sum_{i=1}^{n_j}\begin{pmatrix}1\\ X_{ji}\end{pmatrix}\left(\hat{n}^q_{ji,nn0} + \hat{n}^q_{ji,cn0} + \hat{n}^q_{ji,cc0}\right)\left(Y_{ji} - \eta(X_{ji};\hat{\beta}_{\bullet n},\epsilon^q_{j,\bullet n})\right)
\end{pmatrix}.$$

Now that we have the formula for the observed data information matrix, $I_{\mathbf{n}}$, by the delta method an asymptotically consistent estimate of the variance of $ACE(cc|x)_{\text{cluster}}$ for a fixed value of x is

$$\widehat{Var}(ACE(cc|x)_{\text{cluster}})$$

$$= \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\intercal} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \dfrac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^{\intercal} I_{\mathbf{n}}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \dfrac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc0}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc0}})^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{\mathsf{X}e^{\mathsf{X}\hat{\beta}_{cc1}}}{(1+e^{\mathsf{X}\hat{\beta}_{cc1}})^2} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$