ABSTRACT:

Most investigations in the social and health sciences aim to understand the directional or causal relationship between a treatment or risk factor and outcome. Furthermore, given the multitude of pathways through which the treatment or risk factor may affect the outcome, there is also interest in decomposing the effect of a treatment of risk factor into “direct” and “mediated” effects. For example, child socioeconomic status (risk factor) may have a direct effect on the risk of death (outcome) and an effect that may be mediated through the adulthood socioeconomic status (mediator). Building on the potential outcome framework for causal inference, we develop a Bayesian approach for estimating direct and mediated effects in the context of a dichotomous mediator and dichotomous outcome, which is challenging as many parameters cannot be fully identified. We first define the principal strata corresponding the joint distribution of the observed and counterfactual values of the mediator, and define associate, dissociative, and mediated effects as functions of the differences in the mean outcome under differing treatment assignments within the principal strata. We then develop the likelihood properties and calculate nonparametric bounds of these causal effects assuming randomization of treatment assignment. Since likelihood theory is not well-developed for non-identifiable
parameters, we consider a Bayesian approach which allows the direct and mediated effects to be expressed in terms of the posterior distribution of the population parameters of interest. This range can be reduced by making further assumptions about the parameters that can be encoded in prior distribution assumptions. We perform sensitivity analysis by using several prior distributions that make weaker assumptions than monotonicity or the exclusion restriction. A simulation study investigates the repeated sampling properties of the Bayesian inference on the direct and mediated effects. We also consider an application that explores the mediating effects of adult poverty on the relationship between childhood poverty and risk of death.

KEY WORDS: Direct effect, mediated effect, monotonicity, poverty, mortality.
1 Introduction

Social and health scientists are often interested in understanding how the effect of a risk factor or exposure $Z$ on outcome $Y$ may be mediated through a third factor $D$. For example, children born into poverty may have their lifespan shortened through a variety of mechanisms. One of them might be that childhood poverty causes adult poverty, which in turn leads to reduced lifespan via poor (adult) health care, increased stress, and other factors (Backlund et al. 1996; Wilkerson 1992), so that adult poverty is a mediator between childhood poverty and risk of death.

Alternatively, there may be effects of childhood poverty, for example, poor childhood health care, health knowledge, attitudes and behaviors, and a variety of other impacts that lead directly to reduced lifespan irrespective of adult poverty status (Davey Smith et al. 1998, Kauhanen et al. 2006). The concept of direct and mediated effects is often used in the vaccine literature (Halloran and Struchiner 1995, Haber 1999), where vaccines can reduce the risk of contracting a disease through stimulation of subject’s immune system, or directly affect risk of infection through the “herd effect,” the slowing or stopping of a disease’s movement through a population with an increased overall immune response. Direct and mediated effects are also closely related to issue of inference with a surrogate marker (Prentice 1989; Taylor, Wang, and Thiebaut 2005), where we might anticipate that a good surrogate outcome serves as a mediator of treatment effect, leaving little effect of the treatment to directly impact the true outcome of interest through other channels.

Regression methods for investigating mediation were outlined by Baron and Kenny (1986). They suggest fitting linear models to the data of the form

$$E(Y | Z = z) = \alpha_1 + \beta_1 z$$

$$E(D | Z = z) = \alpha_2 + \beta_2 z$$

$$E(Y | D = d, Z = z) = \alpha_3 + \beta_3 z + \gamma d$$
Mediation is evaluated by considering whether or not there is a significant marginal association between the exposure and outcome ($\beta_1 \neq 0$), whether or not there is a significant association between the exposure and the mediator and between the outcome and the mediator after adjusting for the exposure ($\beta_2 \neq 0$ and $\gamma \neq 0$), and whether or not the direct effect of the exposure on the outcome is smaller in magnitude than the total effect ($|\beta_3| < |\beta_1|$). If all of these conditions are met, then $D$ is said to mediate the effect of $Z$ on $Y$. A special case of mediation occurs when $D$ is considered to be a “surrogate outcome” standing in for the true outcome of interest $Y$: Prentice (1989) discusses a model similar to that of Baron and Kenny, where the definition of a perfect surrogate assumes that $\beta_3 = 0$ after adjusting for $D$. Freedman et al. (1992) and Wang and Taylor (2002) consider situations in which partial surrogacy takes place and define measures of the degree to which these surrogate markers can replace the true outcomes, i.e., the degree to which these surrogate markers mediate the relationship between the treatment and outcome.

A shortcoming of these approaches is that they condition on a post-randomization variable, the observed mediator $D = d$ after the assignment of $Z$. Hence the effect of $Z$ on $Y$ after adjusting for $D$ can no longer be interpreted causally, even if $Z$ is randomly assigned (or can be treated as randomly assigned after adjusting for observed covariates $X$) (Rosenbaum 1984). To get around this, Robins and Greenland (1992) define direct and indirect effects in terms of potential outcomes (Rubin 1974). They consider the set of potential outcomes to include the value of the outcome under each of the possible values of the exposure and mediator, and allow the set of “potential observables” to include values of the mediator under each of the potential exposure assignments. They define a “prescriptive” direct effect as the expected value of the difference in the potential outcomes under different treatment assignments when the value of the mediator is held constant, and an associated prescriptive indirect effect as the expected difference in the total effect (expected difference in potential outcomes under different treatment assignments averaged...
over all values of the mediator) and the prescriptive direct effect. In a setting of dichotomous exposures, mediators, and outcomes, assuming that the exposure is randomized and never improves the value of the mediator and that the effect of the potential exposure and the potential mediator on the outcome do not interact, the direct effect of the exposure on the outcome is the proportion of the population in which the exposure causes the outcome regardless of the potential distribution of the mediator, and the indirect effect is the proportion of the population in which the exposure causes the mediator and the mediator causes the outcome.

Rubin (2004) argues that, instead of allowing for the mediator and the exposure to both be implicitly assignable and thus for the distribution of the full set of potential outcomes to be the product of the distribution of the mediator under all assignments and the distribution of the outcome under all assignments of the mediator and the exposure, that inference should focus on the distribution of the potential outcomes conditional on the distribution of the mediator under all assignments (the “potential mediator”). The values of the mediator under all assignments form pre-randomization “principal strata” (Frangakis and Rubin 2002) within which causal estimators can be obtained. Contrasts in the potential outcomes within strata where the values of the mediator are constant provide an estimate of the direct effect of treatment, while contrasts in the potential outcomes within strata where the values of the mediator change provide an estimate of the mediated effect of treatment. Gallop et al. (2009) and Zhang et al. (2009) have developed this suggestion in the context of a continuous outcome assumed to be normally distributed. Here we assume both a dichotomous mediator and a dichotomous outcome, which provides a challenging problem of identifiability, as we discuss below.

Section 2 defines direct and mediated effects in terms of intent-to-treat effects within the principal strata of the mediators. Section 3 considers the structure of the likelihood and develops Bayesian estimation methods that provide posterior distributions of causal estimates of interest under different a priori constraints on the potential mediator distribution. Section 4 considers the
repeated sampling properties of the Bayesian methods and compares the results with those obtained through standard regression methods that condition on mediators observed post-treatment. Section 5 considers a specific data application, namely estimating the mediating effect of adult poverty on the relationship between childhood poverty and risk of death. Section 6 summarizes our findings and suggests future extensions of these methods. Our work provides two new contributions to the causal modeling literature. First, traditional identifiability restrictions such as monotonicity or the exclusion restriction make strong assumptions about the nature of the population. Thus under monotonicity we assume that the exposure either has no effect or a unidirectional effect on the mediator (in the context of non-compliance, this is the “no-defier” assumption that encouragement to treatment never discourages taking treatment). While this might make sense in some settings, a more reasonable assumption might be that some principal strata are more common than others. Here we consider priors that restrict the orderings of the proportions of principal strata rather than assuming they are zero. Second, we define a “mediated effect” that ranges between 0 and 1 in the absence of directional interaction among the treatment effects in the principal strata as the treatment effect ranges from direct to fully mediated.

2 Direct Effect and Mediated Effect Principal Strata

In this manuscript we focus on the special case of a dichotomous exposure $Z$, dichotomous outcome $Y$, and dichotomous mediator $D$. We denote the potential mediator values under each of the exposure assignments by $D(Z)$, and potential outcome values by $Y(Z, D(Z))$. For patients receiving treatment $Z = z$, we only observe $Y(Z = z, D(Z = z))$ and $D(Z = z)$; $Y(Z = 1 - z, D(Z = 1 - z))$ and $D(Z = 1 - z)$ are unobserved. The joint distribution of $Y(Z, D(Z)), D(Z)$ is a 16-cell multinomial distribution given by Table 1:

\[
P(D(0) = d_0, D(1) = d_1, Y(0) = y_0, Y(1) = y_1) = \pi_{ij}\ 
\text{for } i = 1 \text{ if } d_0 = d_1 = 0, \ i = 2 \text{ if }
\]
\(d_0 = 0, d_1 = 1, i = 3\) if \(d_0 = d_1 = 1\), and \(i = 4\) if \(d_0 = 1, d_1 = 0\), and similarly for \(j, y_0,\) and \(y_1\).

The four sets of values that support the distribution of \(D(Z)\) form the four principal strata within which we will make inference about the potential outcomes \(Y(Z, D(Z))\) and \(Y(1 - Z, D(1 - Z))\):

\[
D(0) = D(1) = 0, D(0) = 0, D(1) = 1, D(0) = D(1) = 1, \text{ and } D(0) = 1, D(1) = 0.
\]

We refer to these principal strata as “never mediators,” “concordant mediators,” “always mediators,” and “discordant mediators.”

The overall causal effect (ce) of the exposure is given by the intent-to-treat (ITT) effect: the contrast of the potential outcome under \(Z = 1\) with the potential outcome under \(Z = 0\):

\[
\sum_{D(Z)} E(Y(1, D(1)) - Y(0, D(0))) = E(Y(1) - Y(0)) = (\pi_{2+} + \pi_{4+}) - (\pi_{3+} + \pi_{4+}) = \pi_{2+} - \pi_{4+}.
\]

Our goal is to make inference about the ITT effect within each of the mediation strata.

Expanding the terminology of Frangakis and Rubin (2002) with respect to surrogate measures, we term the contrast between potential outcomes within strata where the exposure changes the mediator associative effects

\[
E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) \neq D(0)) = ((\pi_{22} + \pi_{42}) - (\pi_{24} + \pi_{44})) / (\pi_{2+} + \pi_{4+})
\]

and the contrast between potential outcomes within strata where the exposure has no effect on the mediator disassociative effects

\[
E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) = D(0)) = ((\pi_{12} + \pi_{32}) - (\pi_{14} + \pi_{34})) / (\pi_{1+} + \pi_{3+}).
\]

If the effect of the exposure is entirely direct, i.e., unmediated through \(D\), then

\(D(Z) \perp Y(Z, D(Z)),\) and

\[
E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) \neq D(0)) = E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) = D(0)) = E(Y(1) - Y(0))
\]

thus

\[
((\pi_{22} + \pi_{42}) - (\pi_{24} + \pi_{44})) / (\pi_{2+} + \pi_{4+}) = ((\pi_{12} + \pi_{32}) - (\pi_{14} + \pi_{34})) / (\pi_{1+} + \pi_{3+}) = \pi_{2+} - \pi_{4+}
\]
or \( ae = de = ce \). If the effect of the exposure is entirely mediated through \( D \), i.e., there is no direct effect of \( Z \) on \( Y \), then

\[
E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) = D(0)) = 0; E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) \neq D(0)) = E(Y(1) - Y(0))
\]

and

\[
((\pi_{22} + \pi_{42}) - (\pi_{24} + \pi_{44})) = \pi_{+2} - \pi_{+4}
\]

or \( ae = \frac{\pi_{+2} - \pi_{+4}}{\pi_{+2} + \pi_{+4}}, de = 0 \). Thus we construct a mediated effect measure

\[
me = \frac{ae - (\pi_{+2} - \pi_{+4})}{\frac{\pi_{+2} - \pi_{+4}}{\pi_{+2} + \pi_{+4}} - (\pi_{+2} - \pi_{+4})}
\]

\[
= \frac{\pi_{22} + \pi_{42} - (\pi_{24} + \pi_{44})}{\pi_{+2} - \pi_{+4}} - \frac{\pi_{2+} + \pi_{4+}}{1 - (\pi_{2+} + \pi_{4+})}.
\]

The intuition behind \( me \) is that it will vary from 0 when the effect of treatment \( Z \) is entirely direct (in which case \( ae = \pi_{+2} - \pi_{+4} \)) to 1 when the effect of \( Z \) is entirely mediated through \( D \) (in which case \( \frac{\pi_{22} + \pi_{42} - (\pi_{24} + \pi_{44})}{\pi_{+2} - \pi_{+4}} = 1 \)). Although \( me \) is technically unbounded, we feel that this measure captures the concept of mediation and direct effect in most settings, since situations where \( me \) is less than 0 or greater than 1 are somewhat pathological. Thus \( me < 0 \) if the ITT effect in the associative strata is smaller than the ITT effect in the disassociative strata – which implies in our poverty example that the effect of childhood poverty on risk of death being stronger when childhood poverty has no impact on adult poverty than when it does – and \( me > 1 \) if disassociative effect is negative – which implies that childhood poverty is \textit{protective} when it has no impact on adult poverty.

Throughout the remainder of this manuscript, we assume that the exposure \( Z \) is assigned at random, so that \( P(Z, D(0), D(1), Y(0, D(0)), Y(1, D(1)) = P(Z) \). Hence the joint distribution of the potential outcomes and potential mediators is independent of treatment assignment, and we avoid the need to directly model the assignment of the treatment \( Z \) (Rubin 1978). We also make the stable unit treatment value assumption (SUTVA) (Rubin 1990), that the assignment of a
Table 1: Joint distribution of counterfactual mediator and outcome.

given subject to treatment $Z_i = z$ is independent of the joint potential outcomes of

$D_j(0), D_j(1), Y_j(0), Y_j(1))$ for $j \neq i$.

### 2.1 Monotonicity Assumption

A common assumption, plausible in many settings, is *monotonicity*: $D(0) \leq D(1)$; this implies no discordant mediators, or $\pi_{4+} = 0$. In the context of the mediating effect of adult poverty on childhood poverty, the no defier assumption implies that no one would experience adult poverty as a consequence of having avoided childhood poverty. If we make the monotonicity assumption for the outcome $Y(0, D(0)) \geq Y(1, D(1))$ as well, that is, there are no “discordant” outcomes (where a subject does worse under a treatment that is designed to help or vice versa), we have $\pi_{4+} = 0$ as well. Hirano et al. (2000) considered a similar model, under the further restriction that either $\pi_{12} = 0$ (exclusion restriction in the never mediators) and/or $\pi_{32} = 0$ (exclusion restriction in the always mediators).

Under the monotonicity assumption the associative effect reduces to

$$ae = E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) \neq D(0)) = \frac{\pi_{22}}{\pi_{2+}}$$
and the disassociative effect to

\[ de = E(Y(1, D(1)) - Y(0, D(0)) \mid D(1) = D(0)) = \frac{\pi_{12} + \pi_{32}}{\pi_{1+} + \pi_{3+}}. \]

The mediated effect measure reduces to

\[ me = \frac{\pi_{22}/\pi_{2+} - \pi_{+2}}{\pi_{2+}/\pi_{+2} - \pi_{+2}} = \frac{\pi_{22}/\pi_{+2} - \pi_{2+}}{1 - \pi_{2+}}. \]

Note that \( ae, de, \) and \( me \) all have an upper bound of 1 under monotonicity, since \( \pi_{22} \leq \pi_{2+}, \) \( \pi_{12} + \pi_{32} \leq \pi_{1+} + \pi_{3+}, \) and the disasociative effect is constrained to be non-negative.

### 3 Inference for Direct and Mediated Effects

#### 3.1 Under the Monotonicity Assumption

We observe \( D(Z = z), Y(Z = z) \) only for the actual treatment assignment \( Z = z. \) Hence the contingency table for observed data is given in Table 2, along with the complete data parameters for each observed data cell. The observed data likelihood is given by

\[
L(\pi; n) = (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})^{n_{00}}(\pi_{13} + \pi_{23})^{n_{01}}(\pi_{31} + \pi_{32})^{n_{10}}(\pi_{33})^{n_{11}} \times \\
\pi_{11}^{n_{00}}(\pi_{12} + \pi_{13})^{n_{01}}(\pi_{21} + \pi_{31})^{n_{10}}(\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33})^{n_{11}}
\]

where \( n_{ij}^{z} \) correspond to the observed cell counts of subjects with \( Z = z, D(z) = i, \) and \( Y(z) = j. \)

Define the observed proportions within each cell as \( p_{ij}^{z} = n_{ij}^{z}/n^{z}. \) Unique MLEs for all marginal row and column percentages are available: \( \hat{\pi}_{+1} = p_{+0}^{1}, \hat{\pi}_{+3} = p_{+1}^{0}, \) and \( \hat{\pi}_{+2} = 1 - \hat{\pi}_{+1} - \hat{\pi}_{+3}; \) similarly \( \hat{\pi}_{1+} = p_{0+}^{1}, \hat{\pi}_{3+} = p_{1+}^{0}, \) and \( \hat{\pi}_{2+} = 1 - \hat{\pi}_{1+} - \hat{\pi}_{3+}. \) Unique MLEs for the upper-left and lower-right cell parameters (\( \pi_{11} \) and \( \pi_{33} \)) can also be identified as \( p_{00}^{11} \) and \( p_{11}^{01}, \) respectively. MLEs for the remaining parameters are not uniquely identified, but exist over a range of values. By considering the constraints imposed by the unique MLEs for sums of the
<table>
<thead>
<tr>
<th>D</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z = 0</td>
<td>( n_{00} )</td>
<td>( n_{01} )</td>
</tr>
<tr>
<td>&amp;</td>
<td>( (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22}) )</td>
<td>( (\pi_{13} + \pi_{23}) )</td>
</tr>
<tr>
<td>1</td>
<td>( n_{10} )</td>
<td>( n_{11} )</td>
</tr>
<tr>
<td>&amp;</td>
<td>( (\pi_{31} + \pi_{32}) )</td>
<td>( (\pi_{33}) )</td>
</tr>
<tr>
<td>Z = 1</td>
<td>( n_{00} )</td>
<td>( n_{01} )</td>
</tr>
<tr>
<td>&amp;</td>
<td>( (\pi_{11}) )</td>
<td>( (\pi_{12} + \pi_{13}) )</td>
</tr>
<tr>
<td>1</td>
<td>( n_{10} )</td>
<td>( n_{11} )</td>
</tr>
<tr>
<td>&amp;</td>
<td>( (\pi_{21} + \pi_{31}) )</td>
<td>( (\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33}) )</td>
</tr>
</tbody>
</table>

Table 2: Observed data table under monotonicity assumption for mediator and outcome.

parameters, MLEs for the remaining six parameters can be identified up to boundaries (Horowitz and Manski, 2000, 2006; Chiba et al. 2007). In particular,

\[
\max(0, p_{00}^0 + p_{11}^1 - 1) \leq \hat{\pi}_{22} \leq \min(p_{00}^0 - p_{00}^1 + \min(0, p_{10}^1 - p_{10}^0), p_{11}^1 - p_{11}^0 + \min(0, p_{01}^1 - p_{01}^0)).
\]

(see Appendix A for derivation). Consequently the boundaries of the MLE for the associative effect are given by

\[
\left( \frac{\max(0, p_{00}^0 + p_{11}^1 - 1)}{1 - p_{+0}^1 - p_{+1}^0}, \frac{\min(p_{00}^0 - p_{00}^1 + \min(0, p_{10}^1 - p_{10}^0), p_{11}^1 - p_{11}^0 + \min(0, p_{01}^1 - p_{01}^0))}{1 - p_{+0}^1 - p_{+1}^0} \right)
\]

for the disassociative effect by

\[
\left( \frac{\max(0, p_{10}^1 - p_{00}^1, p_{01}^0 - p_{00}^1)}{p_{+0}^1 + p_{+1}^0}, \frac{p_{00}^1 - \min(p_{00}^0 - (p_{00}^1 + p_{10}^0), p_{01}^0)}{p_{+0}^1 + p_{+1}^0} \right)
\]

and for the mediated effect by

\[
\left( \frac{\max(0, p_{00}^0 + p_{11}^1 - 1)/(p_{+1}^1 - p_{+1}^0) - (p_{1+}^1 - p_{1+}^0)}{1 - p_{+0}^1 - p_{+1}^0}, \frac{\min(p_{00}^0 - p_{00}^1 + \min(0, p_{10}^1 - p_{10}^0), p_{11}^1 - p_{11}^0 + \min(0, p_{01}^1 - p_{01}^0))/(p_{+1}^1 - p_{+1}^0) - (p_{1+}^1 - p_{1+}^0)}{1 - p_{+0}^1 - p_{+1}^0} \right).
\]
Because the quantities with unique MLEs converge in probability to their true values, the
asymptotic boundaries of the remaining MLEs are given by replacing the point estimates with
their true values. In small samples, the boundaries will be highly variable; as sample size
increases, the boundaries will converge toward their asymptotic limit, with the likelihood
decreasing more rapidly beyond the boundary point. To illustrate this, we consider simulations
from a scenario with equally likely never, concordant, and always mediators in which the effect of
the treatment is entirely through the mediator: \( \pi_{11} = 1/5, \pi_{12} = 0, \pi_{13} = 2/15, \pi_{21} = 0, \)
\( \pi_{22} = 1/3, \pi_{23} = 0, \pi_{31} = 2/15, \pi_{32} = 0, \pi_{33} = 1/5. \) Figure 1 illustrates the profile likelihood for
\( \pi_{22} \) for three samples: \( n = 100, n = 500, n = 2500. \) (The profile likelihood is obtained by fixing
\( \pi_{22} \) at a given value, maximizing the remaining values using an EM algorithm, and computing the
likelihood at the fixed \( \pi_{22} \) and the maximized values of the remaining \( \pi_{ij}; \) see Appendix B for
details.)

Because frequentist theory is poorly developed for situations in which likelihoods are flat,
we turn to Bayesian methods to describe the information available about the associative and
disassociative effects of interest. We obtain simulations from the posterior distribution of \( \pi \) via a
data augmentation algorithm (Tanner and Wong 1987). The complete data is given by the cell
counts \( m^z_{ij} \), where \( z \) indexes the treatment assignment and \( i \) and \( j \) correspond to the indices
previously defined for the counterfactual values of \( D(0), D(1) \) and \( Y(0), Y(1) \) respectively. Under
randomization and SUTVA, we have \( m^z \sim MULTI(n^z; \pi_{11}, ..., \pi_{33}) \), so the data augmentation
step is given by draws from the multinomial distribution:

\[
\begin{align*}
m_{11}^0, m_{12}^0, m_{21}^0, m_{22}^0 & \sim MULTI(n_{00}^0; \frac{\pi_{11}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})}, \frac{\pi_{12}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})}) \\
m_{13}^0, m_{23}^0 & \sim MULTI(n_{01}^0; \frac{\pi_{13}}{(\pi_{13} + \pi_{23})}, \frac{\pi_{23}}{(\pi_{13} + \pi_{23})}) \\
m_{31}^0, m_{32}^0 & \sim MULTI(n_{10}^0; \frac{\pi_{31}}{(\pi_{31} + \pi_{32})}, \frac{\pi_{32}}{(\pi_{31} + \pi_{32})})
\end{align*}
\]
Figure 1: Profile likelihood for $\pi_{22} = 1/3$ for three samples of size 100, 500, and 2500. Asymptotic boundaries for MLE given by dotted vertical lines at $1/15$ and $1/3$. 

\[m_{12}, m_{13} \sim \text{MULTI}(n_{01}, \frac{\pi_{12}}{\pi_{12} + \pi_{13}}, \frac{\pi_{13}}{\pi_{12} + \pi_{13}})\]

\[m_{21}, m_{31} \sim \text{MULTI}(n_{10}, \frac{\pi_{21}}{\pi_{21} + \pi_{31}}, \frac{\pi_{31}}{\pi_{21} + \pi_{31}})\]

\[m_{22}, m_{23}, m_{32}, m_{33} \sim \text{MULTI}(n_{11}, \frac{\pi_{22}}{\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33}}, \frac{\pi_{23}}{\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33}}, \frac{\pi_{32}}{\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33}}, \frac{\pi_{33}}{\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33}})\]

We also have \(n_{33}^0 = n_{11}^0\) and \(m_{00}^1 = n_{00}^1\).

Because the likelihood is flat in a variety of regions of interest in the parameter space, the results will highly sensitive to the choice of the prior distribution, even in large samples. A formal reference prior for cell probabilities in multinomial distributions was provided in Bernardo, in discussion of Kass 1989, as

\[p(\pi_1, \ldots, \pi_m) \propto \prod_{k=1}^{m-1} \left[ \pi_i^{-1/2} \left( 1 - \sum_{j=1}^{k} \pi_j \right)^{-1/2} \right].\]

Using simulation studies we found that the repeated sampling properties of this prior to be less than ideal; instead, using a Dirichlet prior with parameters equal to 1 – the equivalent of a uniform prior on the multinomial parameters under the constraint that the probabilities sum to 1 – gave results that had better asymptotic coverage properties. Consequently we draw \(\pi\) conditional on the previous draw of \(m\) from \(\text{DIRICHLET}(m_{11}^0 + m_{11}^1 + 1, \ldots, m_{33}^0 + m_{33}^1 + 1)\).

### 3.2 Relaxing the Monotonicity Assumption

Allowing for “discordant” mediators and outcomes, the observed data likelihood becomes

\[L(\pi; n) = (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})^{n_{00}^0}(\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})^{n_{01}^0}(\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})^{n_{10}^0}(\pi_{33} + \pi_{34} + \pi_{43} + \pi_{44})^{n_{11}} \times \]

\[(\pi_{11} + \pi_{14} + \pi_{41} + \pi_{44})^{n_{00}^0}(\pi_{12} + \pi_{13} + \pi_{42} + \pi_{43})^{n_{01}^1}(\pi_{21} + \pi_{31} + \pi_{24} + \pi_{34})^{n_{10}^1}(\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33})^{n_{11}^1}\]

Unlike the monotonicity setting, there are no identifiable estimates of any of the parameters governing either the joint distribution or marginal distributions of \(D(0), D(1)\) and
Y(0, D(0)), Y(1, D(1)). There are boundary conditions on the maximum likelihood estimates, however. In particular, the boundary conditions for \(ae\) are

\[
 \left( \begin{array}{c}
 I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \leq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \geq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+u} + \tilde{\pi}_{4+u}) \\
 I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \geq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \leq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+u} + \tilde{\pi}_{4+u}) 
\end{array} \right)
\]

for \(de\) are

\[
 \left( \begin{array}{c}
 I((\tilde{\pi}_{12l} + \tilde{\pi}_{32l}) \leq (\tilde{\pi}_{14u} + \tilde{\pi}_{34u})) (\tilde{\pi}_{1+l} + \tilde{\pi}_{3+l}) + I((\tilde{\pi}_{12l} + \tilde{\pi}_{32l}) \geq (\tilde{\pi}_{14u} + \tilde{\pi}_{34u})) (\tilde{\pi}_{1+u} + \tilde{\pi}_{3+u}) \\
 I((\tilde{\pi}_{12l} + \tilde{\pi}_{32l}) \geq (\tilde{\pi}_{14u} + \tilde{\pi}_{34u})) (\tilde{\pi}_{1+l} + \tilde{\pi}_{3+l}) + I((\tilde{\pi}_{12l} + \tilde{\pi}_{32l}) \leq (\tilde{\pi}_{14u} + \tilde{\pi}_{34u})) (\tilde{\pi}_{1+u} + \tilde{\pi}_{3+u}) 
\end{array} \right)
\]

and for \(me\) are

\[
 \frac{A - B}{1 - I(A \leq B)(\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I(A \geq B)(\tilde{\pi}_{2+u} + \tilde{\pi}_{4+u})}
\]

where

\[
 A = \frac{\pi_{22l} + \pi_{42l} - (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})}{I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \leq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I((\tilde{\pi}_{22l} + \tilde{\pi}_{42l}) \geq (\tilde{\pi}_{24u} + \tilde{\pi}_{44u})) (\tilde{\pi}_{2+u} + \tilde{\pi}_{4+u})}
\]

\[
 B = (\tilde{\pi}_{2+u} + \tilde{\pi}_{4+u})
\]

and the lower and upper MLE limits \(\tilde{\pi}_{i+}\) and \(\tilde{\pi}_{u+}\) for \(\pi_{i+}\), and equivalently \(\tilde{\pi}_{i-}\) and \(\tilde{\pi}_{u-}\) for \(\pi_{i-}\), and \(\tilde{\pi}_{+j}\) and \(\tilde{\pi}_{+j}\) for \(\pi_{+j}\) are provided in Appendix A.

Here the data augmentation step is given by

\[
m_{11}^0, m_{12}^0, m_{21}^0, m_{22}^0 \sim MULTI(n_{00}^0, \frac{\pi_{11}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})}, \frac{\pi_{12}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})}, \frac{\pi_{21}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})}, \frac{\pi_{22}}{(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})})
\]

\[
m_{13}^0, m_{23}^0, m_{14}^0, m_{24}^0 \sim MULTI(n_{01}^0, \frac{\pi_{13}}{(\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})}, \frac{\pi_{23}}{(\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})}, \frac{\pi_{14}}{(\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})}, \frac{\pi_{24}}{(\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})})
\]

\[
m_{31}^0, m_{32}^0, m_{31}^0, m_{42}^0 \sim MULTI(n_{10}^0, \frac{\pi_{31}}{(\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})}, \frac{\pi_{32}}{(\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})}, \frac{\pi_{31}}{(\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})}, \frac{\pi_{32}}{(\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})})
\]
rejecting all draws from the conditional posterior of $\pi$ of "concordant" outcomes be greater than the fraction of "discordant" outcomes. We term this mediators, and that, within all of the principal strata except for the discordant, that the fraction requiring that the fraction of "concordant" mediators be greater than the fraction of "discordant"

We also consider a restricted prior that constrains $\pi$ is that the directionality of the treatment, mediator, and outcome have been "lined up,"

We retain

$\pi \mid m \sim \text{DIRICHLET}(1, \ldots, 1)$

so that

$\pi \mid m \sim \text{DIRICHLET}(m_{11}^0 + m_{11}^1 + 1, \ldots, m_{44}^0 + m_{44}^1 + 1)$.

### 3.3 Stochastic Monotonicity Assumption

We also consider a restricted prior that constrains $\pi_{2+} \geq \pi_{4+}$ and $\pi_{j2} \geq \pi_{j4}$ for $j = 1, 2, 3$, requiring that the fraction of "concordant" mediators be greater than the fraction of "discordant" mediators, and that, within all of the principal strata except for the discordant, that the fraction of "concordant" outcomes be greater than the fraction of "discordant" outcomes. We term this prior “stochastic monotonicity.” The constraint is imposed in the Gibbs sampling process by rejecting all draws from the conditional posterior of $\pi$ that do not meet the constraint. Implicit in this prior is that the directionality of the treatment, mediator, and outcome have been “lined up,”
so that \( Z = 1, \ D = 1 \) and \( Y = 1 \) are consistent with a risk factors and poor outcomes, or protective treatments and good outcomes. As with the monotonicity assumption, this prior implies an upper limit of 1 for the mediated effect \( me \), since negative disassociative effects are no longer possible.

4 Simulation Studies

Although we develop our methods in the Bayesian framework, we consider the repeated sampling properties of the estimators to be important. Hence we conduct a simulation study of the posterior distributions of the proposed mediation effects.

4.1 Under Monotonicity Assumption

We consider four scenarios under the monotonicity assumption: two in which the effect of the treatment is heavily mediated, and two in which the treatment is largely direct. Table 3 shows the joint distribution of counterfactual mediator and outcome for all scenarios. In all scenarios 1/3 of the population belongs to each of the three principal strata, and the causal effect of treatment on outcome \( P(Y(1)) - P(Y(0)) \) is 1/3 as well. The associative effect is .8, the disassociative effect is .1, and the mediated effect is .7 in the heavily mediated scenarios; the associative effect is .4, the disassociative effect is .3 and the mediated effect is .1 in the direct effect scenarios. Scenario (a) corresponds to a situation in which the potential outcomes for the mediator and outcomes are in correspondence: the treatment increases the probability of a “success” for both the mediator and outcome, the never mediators and doomed outcomes where \( Y(0) = Y(1) = 0 \) are highly correlated, and the treatment effect is largely mediated. Scenario (b) corresponds to a case where never mediator and doomed outcomes are uncorrelated despite the treatment being causally associated with the mediator and the treatment effect being largely mediated. Scenario (d) corresponds to a
situation where the causal effect of treatment is largely direct and the mediator is independent of the outcome. Scenario (c) is a case where the causal effect of treatment is direct but the mediator and outcome are correlated through the never and always mediator strata. The sample sizes for each simulation are 1,000 and 10,000; each subject is independently assigned to treatment or control with equal probability. For each simulated dataset, the posterior distribution of $\pi$ was obtained via a Gibbs sampling chain of 250,000 draws after discarding the first 1000.

The results of the simulation are reported in Tables 4 ($n=1,000$) and 5 ($n=10,000$). When mediation is strong and counterfactual correlation is strong (scenario (a)), the standard approach works well and $ae$, $de$, and $me$ are well-estimated. When mediation is strong and counterfactual correlation is weak (scenario (c)), the standard approach incorrectly suggests that little or no mediation is present. In contrast, the counterfactual model suggests that $ae$, $de$, and $me$ are can
no longer be well-estimated, and a single point estimate such as a median is no longer appropriate. The resulting wide intervals for \( ae, de, \) and \( me \) maintain conservative coverage, and the lack of information in the posterior is a signal that strong mediation can neither be concluded or rejected from the available data: see Figure 2 for an example posterior of \( me \) under scenarios (a) and (b). Similarly, when mediation and counterfactual correlation are both weak (scenario (d)), lack of mediation can be correctly assessed by the Baron and Kenney method. When mediation is weak and counterfactual correlation is strong (scenario (c)), the Baron and Kenney method can suggest mediation is present; in contrast, the posterior intervals for \( ae, de, \) and \( me \) widen to reflect the uncertainty in the mediated effect that can be estimated from the observed data as the counterfactual correlation changes.

Point estimates have somewhat unappealing properties in this setting due to lack of identifiability. Using the median leads results that center around the middle of the asymptotic MLE interval as sample size increases. However, the resulting 95% credible intervals have lengths that can be up to 30% shorter on average than the lengths of the asymptotic MLE intervals, while maintaining good or even substantially conservative coverage. Indeed, when the true effect is not near the boundary of the asymptotic MLE interval, the coverage of the 95% credible intervals is close to 100%, reflecting the fact that the posterior will reflect the prior in the range of this interval, and thus will include the true value with probability 0 or 100, depending on the distribution of the prior conditional on the range of the asymptotic MLE interval. Values nearer the edge of asymptotic MLE interval may have substantial undercoverage in small samples, although this coverage will also approach 0% or 100% depending on the conditional prior. In the simulations considered here all coverages will reach 100% will sufficiently large sample sizes.
Table 4: Simulation results: Median, posterior interval (PI) length, and coverage of 95% PI for associative effect \((ae)\), disassociative effect \((de)\), and mediated effect \((me)\), along with length of MLE interval, and the MLE and associated confidence interval (CI) of log-odds of unadjusted treatment effect and treatment effect adjusted for mediator using (1.1). Means for 200 simulations. Sample size of 1000. Description of scenarios given in Table 3.
Table 5: Simulation results: Median, posterior interval (PI) length, and coverage of 95% PI for associative effect (*ae*), disassociative effect (*de*), and mediated effect (*me*), along with length of MLE interval, and the MLE and associated confidence interval (CI) of log-odds of unadjusted treatment effect and treatment effect adjusted for mediator using (1.1). Means for 200 simulations. Sample size of 10000. Description of scenarios given in Table 3.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Heavily Mediated Scenarios</th>
<th>Largely Direct Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td><em>ae</em> (asy. boundary)</td>
<td>.8 (.7-1)</td>
<td>.8 (0-.9)</td>
</tr>
<tr>
<td>Median</td>
<td>.84</td>
<td>.44</td>
</tr>
<tr>
<td>PI Length</td>
<td>.23</td>
<td>.79</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td><em>de</em> (asy. boundary)</td>
<td>.1 (0-.15)</td>
<td>.1 (.05-.5)</td>
</tr>
<tr>
<td>Median</td>
<td>.08</td>
<td>.28</td>
</tr>
<tr>
<td>PI Length</td>
<td>.11</td>
<td>.40</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td><em>me</em> (asy. boundary)</td>
<td>.7 (.55-1)</td>
<td>.7 (-.5-.85)</td>
</tr>
<tr>
<td>Median</td>
<td>.77</td>
<td>.16</td>
</tr>
<tr>
<td>PI Length</td>
<td>.34</td>
<td>1.19</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>Unadj. Treatment (CI Length)</td>
<td>1.38 (.08)</td>
<td>1.38 (.08)</td>
</tr>
<tr>
<td>Adj. Treatment (CI Length)</td>
<td>.77 (.13)</td>
<td>1.17 (.09)</td>
</tr>
</tbody>
</table>
Figure 2: Example posterior distribution of mediated effect under scenarios (a), (b), (c), and (d).

True values shown with solid line.
4.2 Relaxing the Monotonicity Assumption

We consider four scenarios that relax the monotonicity assumption, paralleling the scenarios under monotonicity: (e) and (f) are heavily mediated, and (g) and (h) are largely direct. Table 6 shows the joint distribution of counterfactual mediator and outcome for all scenarios. In the heavy mediation scenarios (e) and (f) 40% of the population belongs to the concordant mediator stratum, 25% of the population to the always and never mediator stratum, and 10% of the population to the discordant mediator stratum. The causal effect of treatment on outcome is .2667. In the direct effect scenarios (g) and (h) 33% of the population belongs to the concordant mediator stratum, 25% of the population to the always and never mediator stratum, and 17% of the population to the discordant mediator stratum. The causal effect of treatment on outcome is .1584. Each subject is independently assigned treatment or control with equal probability. For each simulated dataset, the posterior distribution of \( \pi \) was obtained via a Gibbs sampling chain of 250,000 draws after discarding the first 1000. We consider a flat Dirichlet prior (Table 7) and a stochastic monotonicity prior where the fraction of concordant mediators is greater than the fraction of discordant mediators, and the fraction of “concordant” outcomes is greater than the fraction of “discordant” outcomes except in the defier stratum (Table 8). Our focus in this section is to compare different priors, so as to avoid additional complexity we consider a single sample size of 10,000.

The results broadly parallel the results under the monotonicity assumption. The interval widths are now substantially wider, reflecting the reduced information available when fewer constraints are provided for either the mediator or outcome variable. However, the posterior interval lengths are much shorter than the asymptotic MLE boundaries (in the case of dissociative effects these lengths are infinite); constraining the prior for \( \pi \) further reduces these posterior interval lengths by 15%-30%. Situations where standard mediation approaches assess a strong
<table>
<thead>
<tr>
<th></th>
<th>$Y(0, D(0)), Y(1, D(1))$</th>
<th>$Y(0, D(0)), Y(1, D(1))$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heavily Mediated</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0)</td>
<td>.1750 .0333 .0250 .0167</td>
<td>.1000 .0375 .1000 .0125</td>
</tr>
<tr>
<td>$D(0), D(1)$</td>
<td>(0,1)</td>
<td></td>
</tr>
<tr>
<td>(0,1)</td>
<td>.0200 .3267 .0200 .0333</td>
<td>.0400 .2666 .0400 .0536</td>
</tr>
<tr>
<td>(1,1)</td>
<td>.0250 .0333 .1750 .0167</td>
<td>.1000 .0375 .1000 .0125</td>
</tr>
<tr>
<td>(1,0)</td>
<td>.0100 .0100 .0100 .0700</td>
<td>.0100 .0116 .0700 .0083</td>
</tr>
<tr>
<td><strong>Direct Effect</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0)</td>
<td>.1125 .0875 .0125 .0375</td>
<td>.0625 .0875 .0625 .0375</td>
</tr>
<tr>
<td>$D(0), D(1)$</td>
<td>(0,1)</td>
<td></td>
</tr>
<tr>
<td>(0,1)</td>
<td>.0833 .1167 .0833 .0500</td>
<td>.0833 .1167 .0833 .0500</td>
</tr>
<tr>
<td>(1,1)</td>
<td>.0125 .0875 .1125 .0375</td>
<td>.0625 .0750 .0625 .0500</td>
</tr>
<tr>
<td>(1,0)</td>
<td>.0417 .0500 .0417 .0333</td>
<td>.0417 .0500 .0417 .0333</td>
</tr>
</tbody>
</table>

Table 6: Simulation scenarios without monotonicity assumption.
degree of mediation are estimated with narrower credible intervals than situations where the standard mediation approach fails. When effects are largely direct credible intervals are wider. Posterior medians for $ae$ and $de$ tend to be centered around the true values moreso than when monotonicity can be assumed, although the interval widths remain quite wide. Point estimates of $me$ using the posterior median were more accurate for three of the four scenarios when the stochastic monotonicity prior was assumed; neither prior produced an accurate estimate when the treatment was largely mediated but the joint counterfactual distribution of $D$ and $Y$ was small (scenario (f)). The flat likelihood means that asymptotically credible interval coverage will again approach 0 or 100%, although again all intervals approach 100% in this set of simulations.

5 Application: Mediating Effects of Adult Poverty on Risk of Death Due to Childhood Poverty

The Alameda County Study is a stratified random sample survey of households living in Alameda County in California (Breslow and Kaplan, 1965). The purpose of the survey was to explore the influence of health practices and social relationships on the physical and mental health of a representative sample of the Alameda County population. Information was obtained for 6,928 respondents covering chronic health conditions, health behaviors, social involvement, and psychological characteristics. Questions were asked on marital and life satisfaction, parenting, physical activities, employment, and childhood experiences. Demographic variables on age, race, height, weight, education, income, and religion are also included. In particular, poverty during childhood and poverty at the time of the 1965 interview are ascertained. Respondents were followed and survival status noted for 3,352 respondents in 2000. Survival status by childhood poverty status and adult poverty status is shown in Table 9: 28% of children not in poverty were adults in poverty, and 44% of children in poverty were also adults in poverty.
<table>
<thead>
<tr>
<th></th>
<th>Heavily Mediated Scenarios</th>
<th>Largely Direct Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(e)</td>
<td>(f)</td>
</tr>
<tr>
<td>ae (asy. boundary)</td>
<td>.47 (-.86-2.11)</td>
<td>.43 (-1.12-2.01)</td>
</tr>
<tr>
<td>Median</td>
<td>.44</td>
<td>.33</td>
</tr>
<tr>
<td>PI Length</td>
<td>.42</td>
<td>.61</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>(g)</td>
<td>(h)</td>
</tr>
<tr>
<td>de (asy. boundary)</td>
<td>.07 (∞)</td>
<td>.1 (∞)</td>
</tr>
<tr>
<td>Median</td>
<td>-.02</td>
<td>.20</td>
</tr>
<tr>
<td>PI Length</td>
<td>.50</td>
<td>.63</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>me (asy. boundary)</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>1.09</td>
<td>.26</td>
</tr>
<tr>
<td>PI Length</td>
<td>1.90</td>
<td>2.36</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Unadj. Treatment (CI Length)</td>
<td>1.09 (.08)</td>
<td>1.10 (.08)</td>
</tr>
<tr>
<td>Adj. Treatment (CI Length)</td>
<td>.52 (.11)</td>
<td>.94 (.08)</td>
</tr>
</tbody>
</table>

Table 7: Simulation results: Median, posterior interval (PI) length, and coverage of 95% PI for associative effect \(ae\), disassociative effect \(de\), and mediated effect \(me\), along with length of MLE interval, and the MLE and associated confidence interval (CI) of log-odds of unadjusted treatment effect and treatment effect adjusted for mediator using (1.1). Means for 200 simulations. Sample size of 10,000. Description of scenarios given in Table 6.
<table>
<thead>
<tr>
<th></th>
<th>Heavily Mediated Scenarios</th>
<th>Largely Direct Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(e)</td>
<td>(f)</td>
</tr>
<tr>
<td>ae (asy. boundary)</td>
<td>.47 (-.86-2.11)</td>
<td>.43 (-1.12-2.01)</td>
</tr>
<tr>
<td></td>
<td>.37</td>
<td>.27</td>
</tr>
<tr>
<td>PI Length</td>
<td>.35</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de (asy. boundary)</td>
<td>.07 (∞)</td>
<td>.1 (∞)</td>
</tr>
<tr>
<td></td>
<td>.09</td>
<td>.26</td>
</tr>
<tr>
<td>PI Length</td>
<td>.31</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>me (asy. boundary)</td>
<td>.75 (-3.10-3.27)</td>
<td>.62 (-3.42-2.96)</td>
</tr>
<tr>
<td></td>
<td>.65</td>
<td>.03</td>
</tr>
<tr>
<td>PI Length</td>
<td>1.18</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Unadj. Treatment (CI Length)</td>
<td>1.09 (.08)</td>
<td>1.10 (.08)</td>
</tr>
<tr>
<td>Adj. Treatment (CI Length)</td>
<td>.52 (.11)</td>
<td>.94 (.08)</td>
</tr>
</tbody>
</table>

Table 8: Simulation results: Median, posterior interval (PI) length, and coverage of 95% PI for associative effect (ae), disassociative effect (de), and mediated effect (me), along with length of MLE interval, and the MLE and associated confidence interval (CI) of log-odds of unadjusted treatment effect and treatment effect adjusted for mediator using (1.1). Means for 200 simulations. Sample size of 10,000. Description of scenarios given in Table 6.
<table>
<thead>
<tr>
<th>Survival Status</th>
<th>Alive</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Not In Childhood Poverty</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not in Poverty as Adult</td>
<td>1329</td>
<td>254</td>
</tr>
<tr>
<td>In Poverty as Adult</td>
<td>507</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>1836</td>
<td>376</td>
</tr>
<tr>
<td><strong>In Childhood Poverty</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not in Poverty as Adult</td>
<td>504</td>
<td>130</td>
</tr>
<tr>
<td>In Poverty as Adult</td>
<td>369</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>873</td>
<td>267</td>
</tr>
</tbody>
</table>

Table 9: Childhood poverty status, adult poverty status, and survival status of Alameda County Study subjects.

It could be argued that childhood poverty is a largely randomized variable in that children do not choose their poverty status but are in some sense “randomly” assigned at birth; alternatively an analysis could be conducted that controlled for confounders such as size of family, parent’s marital status, or other factors associated with childhood poverty that one might wish to separate from the pure effect of poverty by using a preliminary propensity score adjustment (Rosenbaum and Rubin 1983). Here we use propensity scores to restore balance with respect to gender, age, and race between those in poverty and those not in poverty during childhood. We include a linear and quadratic term for age and a dummy variable for race (white, African-American, and other) to account for the fact that older persons were more likely to experience childhood poverty than younger persons, and African-Americans more likely and those of other races less likely to experience childhood poverty. Because of the extraordinary imbalance with respect to race among those in childhood poverty, African-Americans remain more likely to be in childhood poverty than whites, although the difference is substantially reduced (see Table
First, we conduct an analysis of the form that Baron and Kenny (1986) proposed. Children in poverty are more likely to be in poverty as adults than children not in poverty (OR=2.00, 95% CI=1.72-2.32), showing an association between the exposure and potential mediator. The unadjusted odds ratio of death for persons in childhood poverty is 1.50 (95% CI 1.26-1.79); adjusting for adult poverty reduces this association only slightly (OR=1.43, 95% CI=1.20-1.71), suggesting that most of the effect of childhood poverty on risk of death is direct, and not mediated by the increased risk of being in adult poverty. Adjusting for gender, age and race reduces the overall effect of childhood poverty on risk of death (OR=1.20, 95%CI=0.99-1.45), and suggests a partial degree of mediation through adult poverty among this remaining effect (OR=1.13, 95%=.93-1.37).

Next, we conduct an analysis considering the associative, disassociative, and mediated effects unconstrained, under the stochastic monotonicity assumption, and under the deterministic monotonicity assumption. Table 11 shows the 5th, 50th, and 95th percentiles for the associative, disassociative, and mediated effects, along with the fraction of the population that is estimated to be in each of the mediator principal strata, unadjusted and adjusted for age and race using the propensity scores described above. The propensity-score-adjusted analysis was conducted by

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>-.045(.073)</td>
<td>.010(.076)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>.036(.005)</td>
<td>-.009(.009)</td>
</tr>
<tr>
<td>African-American (vs. white)</td>
<td>.930(.117)</td>
<td>.552(.146)</td>
</tr>
<tr>
<td>Other (vs. white)</td>
<td>-.374(.191)</td>
<td>.023(.200)</td>
</tr>
</tbody>
</table>

Table 10: Log OR of being in childhood poverty, unadjusted and adjusted for propensity score quintile (standard error in parenthesis).
stratifying the data by propensity score quintile and running separate MCMC chains within each stratum. A draw from the posterior of $\pi$ is obtained as the weighted average of draws from each of the five propensity strata, weighted in proportion to the fraction of the sample contained in each (approximate) quintile.

Allowing for non-monotonicity suggests that about 40% of the population (95% CI 35-45%) is immune to adult poverty (never mediators), 30% (95% CI 26-36%) are protected against adult poverty by not experiencing childhood poverty (concordant mediators), 15% (95% CI 11-19%) are doomed to adult poverty (always mediators), and 15% (95% CI 10-18%) experience adult poverty only if they do not experience childhood poverty (discordant mediators). The associative effect and disassociative effects are approximately equal, with no strong evidence of mediation effects, although the possibility cannot be discounted under the unconstrained prior. Constraining the prior has little effect on the posterior median of the associative effect, although the intervals are substantially reduced; however the disassociative effect is estimated to be larger, suggesting rather counterintuitively that the effect of childhood poverty on survival is actually stronger when there is no impact on adult poverty than when there is. Adjusting for age and race via the propensity score analysis tends to move the posterior medians toward zero – reflecting that part of the childhood poverty effect is confounded with the age and race of the respondents – and reduces the width of the posterior intervals. For comparison purposes, the unadjusted overall effect of death on childhood poverty is 6.4 percentage points, and the age/race/sex adjusted effect is 2.6 percentage points.

Assuming monotonicity suggests the fraction of never mediators in 54% (95% CI 51-56%), of concordant mediators is 18% (95% CI 15-21%), and of always mediators is 28% (95% CI 27-30%). The disassociative effect is centered near the point estimate for the overall effect of childhood poverty. The associative effect is generally centered near 0, although the possibility of a substantial associative effect suggesting mediation through adult poverty cannot be entirely ruled
out. Constraining the prior has little effect on the posterior medians, but does shrink the posterior intervals for the mediation effects substantially. Figure 3 shows the posterior distributions of $ae$, $de$, and $me$ adjusting for age, race, and gender, under the three prior assumptions considered.

6 Discussion

Standard regression approaches such as Baron and Kenney (1986) to mediation lack causal interpretation due to potential unobserved confounding even when treatment is randomized, because mediator is observed post-randomization. Use of principal strata defined using the counterfactual distribution of the mediators creates a conceptual pre-randomization variable. In particular, a disassociative effect of treatment can be estimated as the intent-to-treat (ITT) effect among subjects for whom the mediator does not change under different treatment assignments, and similarly an associative effect can be estimated as the ITT effect among subjects for whom the mediator does change under different treatment assignments. A mediated effect can then be constructed by considering the value of the disassociative and associative effects when the overall treatment effect is entirely direct versus completely mediated.

In the setting we consider here – dichotomous mediators and treatments – lack of identifiability suggests use of Bayesian inference. Posterior distributions of ITT effects within principal strata are informed by the data, since boundary conditions are imposed on the counterfactual distribution. In general, principal stratum inference is consistent with standard regression analysis when counterfactual correlation between mediator and outcome is large(small) when mediation is present(absent). Principal stratum inference analysis protects against inappropriate inference when counterfactual correlation is small(large) when mediation is present(absent). The Bayesian approach also allow us to incorporate constraints such as monotonicity ($D(0) \leq D(1)$) or a relaxed stochastic monotonicity.
<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th></th>
<th></th>
<th>Adjusted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5th</td>
<td>50th</td>
<td>95th</td>
<td>5th</td>
<td>50th</td>
<td>95th</td>
</tr>
<tr>
<td>Unconstrained (MLE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ae$ (-1.07-1.46)</td>
<td>-.094</td>
<td>.076</td>
<td>.237</td>
<td>-.044</td>
<td>.035</td>
<td>.114</td>
</tr>
<tr>
<td>$de$ (-.62-.86)</td>
<td>-.074</td>
<td>.054</td>
<td>.179</td>
<td>-.030</td>
<td>.034</td>
<td>.098</td>
</tr>
<tr>
<td>$me$ (-4.07-4.20)</td>
<td>-1.857</td>
<td>.159</td>
<td>2.28</td>
<td>-2.32</td>
<td>.018</td>
<td>2.31</td>
</tr>
<tr>
<td>Never mediators (.271-.556)</td>
<td>.329</td>
<td>.416</td>
<td>.502</td>
<td>.353</td>
<td>.402</td>
<td>.451</td>
</tr>
<tr>
<td>Concordant mediators (.159-.444)</td>
<td>.211</td>
<td>.298</td>
<td>.384</td>
<td>.258</td>
<td>.308</td>
<td>.356</td>
</tr>
<tr>
<td>Always mediators (0-.284)</td>
<td>.062</td>
<td>.145</td>
<td>.229</td>
<td>.106</td>
<td>.148</td>
<td>.192</td>
</tr>
<tr>
<td>Discordant mediators (0-.284)</td>
<td>.057</td>
<td>.139</td>
<td>.225</td>
<td>.098</td>
<td>.141</td>
<td>.185</td>
</tr>
<tr>
<td>Stochastic Monotonicity (MLE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ae$ (-1.07-1.46)</td>
<td>-.026</td>
<td>.057</td>
<td>.153</td>
<td>-.008</td>
<td>.035</td>
<td>.082</td>
</tr>
<tr>
<td>$de$ (-.62-.86)</td>
<td>.021</td>
<td>.076</td>
<td>.161</td>
<td>.046</td>
<td>.076</td>
<td>.116</td>
</tr>
<tr>
<td>$me$ (-4.07-4.20)</td>
<td>-1.346</td>
<td>-.117</td>
<td>.672</td>
<td>-1.066</td>
<td>-.331</td>
<td>.172</td>
</tr>
<tr>
<td>Never mediators (.271-.556)</td>
<td>.320</td>
<td>.413</td>
<td>.504</td>
<td>.347</td>
<td>.397</td>
<td>.447</td>
</tr>
<tr>
<td>Concordant mediators (.159-.444)</td>
<td>.209</td>
<td>.302</td>
<td>.393</td>
<td>.262</td>
<td>.313</td>
<td>.362</td>
</tr>
<tr>
<td>Always mediators (0-.284)</td>
<td>.055</td>
<td>.143</td>
<td>.232</td>
<td>.101</td>
<td>.145</td>
<td>.191</td>
</tr>
<tr>
<td>Discordant mediators (0-.284)</td>
<td>.055</td>
<td>.143</td>
<td>.232</td>
<td>.100</td>
<td>.145</td>
<td>.190</td>
</tr>
<tr>
<td>Monotonicity (MLE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ae$ (0-1)</td>
<td>.010</td>
<td>.125</td>
<td>.330</td>
<td>.055</td>
<td>.117</td>
<td>.201</td>
</tr>
<tr>
<td>$de$ (0-.069)</td>
<td>.019</td>
<td>.056</td>
<td>.093</td>
<td>.034</td>
<td>.054</td>
<td>.077</td>
</tr>
<tr>
<td>$me$ (-.190-1)</td>
<td>-.161</td>
<td>.159</td>
<td>.690</td>
<td>-.024</td>
<td>.175</td>
<td>.406</td>
</tr>
<tr>
<td>Never mediators (.556)</td>
<td>.531</td>
<td>.555</td>
<td>.579</td>
<td>.513</td>
<td>.538</td>
<td>.562</td>
</tr>
<tr>
<td>Concordant mediators (.160)</td>
<td>.132</td>
<td>.161</td>
<td>.189</td>
<td>.151</td>
<td>.179</td>
<td>.209</td>
</tr>
<tr>
<td>Always mediators (.284)</td>
<td>.269</td>
<td>.284</td>
<td>.300</td>
<td>.266</td>
<td>.283</td>
<td>.299</td>
</tr>
</tbody>
</table>

Table 11: Posterior 5th, 50th, and 95th percentiles for associative, disassociative, and mediated effects and for proportion of population in principal strata mediator classes: unconstrained, under the stochastic monotonicity assumption, and under the deterministic monotonicity assumption. Unadjusted and adjusted for age and race.
Figure 3: Posterior distribution of associative ($ae$), disassociative ($de$), and mediated effect ($me$) of childhood poverty of risk of death mediated through adult poverty, adjusted for age, race, and gender using propensity scores. SM=Stochastic monotonicity assumption; Mono=monotonicity assumption.
\(P(D(0) = 1, D(1) = 0) \leq P(D(0) = 0, D(1) = 1))\). These constraints may be used in a single analysis if reasonable, or incorporated in a set of sensitivity analysis as was done in the example in this manuscript.

The methods developed here assume unconfounded treatment assignment \(Z\). This assumption is very strong in many observational studies, although occasionally treatments of interest might appear in the form of instrumental variables, such as changes in laws that might effect mediating behaviors but would be independent of the joint distribution of the potential outcomes. In general we would propose use of propensity scores to balance on observed covariates, as we did when assessing the mediating effect of adult poverty on this risk of death due to childhood poverty.

Many extensions of this work are possible. A variety of different prior constraints could be considered: for example, we might retain monotonicity for the outcome but relax it in some fashion for the mediator, or vice versa. Baseline covariates that allow prediction of principal stratification status can be useful in sharpening inference. In the non-compliance setting, where much of the mediation analysis using principal strata has focused, few practical predictors of compliance have been found. In more general applications, such as the one considered here, searches for predictors of principal stratification status may be more fruitful.

ACKNOWLEDGEMENTS

The author would like to thank Jeremy Taylor, Thomas Ten Have, Dylan Small, and Marshall Joffe for their helpful comments. This research was supported by National Institute of Mental Health Grant R01MH-078016 and National Cancer Institute Grant R01CA-129102.
A Appendix A: Derivation of Boundaries for Associative and Disassociative Effects

A.1 Under Monotonicity.

Consider the observed data likelihood given by (3.2) under the assumption of monotonicity for the mediator and outcome. Unique MLEs for \( \pi_{11} \) and \( \pi_{33} \) are given by \( \hat{\pi}_{11} = p_{00}^1 \) and \( \hat{\pi}_{33} = p_{11}^0 \), with the remaining MLEs identified only up to sums:

\[
\hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} = p_{00}^0 \tag{A.1}
\]

\[
\hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} = p_{11}^1 \tag{A.2}
\]

subject to \( \sum_{i=1}^3 \sum_{j=1}^3 \hat{\pi}_{ij} = 1 \).

From (A.1), we have \( p_{00}^0 - p_{00}^1 = \hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} - \hat{\pi}_{11} = \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} \geq \hat{\pi}_{22} \). From (A.2), we have \( p_{11}^1 - p_{11}^0 = \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} - \hat{\pi}_{33} = \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} \geq \hat{\pi}_{22} \). Putting (A.1) and (A.2) together, we have

\[
p_{00}^0 + p_{11}^1 - 1 = \hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} + \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} - 1 = \hat{\pi}_{22} - \hat{\pi}_{13} - \hat{\pi}_{31} \leq \hat{\pi}_{22}.
\]

We also have from \( p_{01}^0 - p_{10}^1 = \hat{\pi}_{31} + \hat{\pi}_{32} - (\hat{\pi}_{21} + \hat{\pi}_{31}) = \hat{\pi}_{32} - \hat{\pi}_{21} \) that

\[
p_{00}^0 - p_{00}^1 + (p_{10}^1 - p_{10}^0) = \hat{\pi}_{12} + \hat{\pi}_{22} + \hat{\pi}_{32} = \hat{\pi}_{+2} \leq \hat{\pi}_{22}, \text{ and similarly } p_{11}^1 - p_{11}^0 + (p_{01}^0 - p_{01}^0) \leq \hat{\pi}_{22}.
\]

Thus

\[
\max(0, p_{00}^0 + p_{11}^1 - 1) \leq \hat{\pi}_{22} \leq \min(p_{00}^0 - p_{00}^1 + \min(0, p_{10}^1 - p_{10}^0), p_{11}^1 - p_{11}^0 + \min(0, p_{01}^0 - p_{01}^0)).
\]

Because \( \hat{\pi}_{12} + \hat{\pi}_{32} = \hat{\pi}_{+2} \), we have

\[
\hat{\pi}_{+2} - \min(p_{00}^0 - p_{00}^1, p_{11}^1 - p_{11}^0) \leq \hat{\pi}_{12} + \hat{\pi}_{32} \leq \hat{\pi}_{+2} - \max(0, p_{00}^0 + p_{11}^1 - 1).
\]

Using \( \hat{\pi}_{+2} = 1 - p_{+0}^1 - p_{+1}^0 = p_{00}^0 + p_{11}^0 - (p_{00}^1 + p_{11}^1), \) we have

\[
p_{10}^0 + p_{00}^0 - (p_{00}^1 + p_{11}^1) - \min(p_{00}^0 - p_{00}^1, p_{11}^1 - p_{11}^0) \leq \hat{\pi}_{12} + \hat{\pi}_{32} \leq p_{10}^0 + p_{00}^0 - (p_{00}^1 + p_{11}^1) - \max(0, p_{00}^0 + p_{11}^1 - 1).
\]
or
\[
\max(0, p_{10}^0 - p_{00}^1, p_{01}^1 - p_{10}^0) \leq \hat{\pi}_{12} + \hat{\pi}_{32} \leq p_{10}^0 + \min(p_{00}^0, p_{10}^1, p_{01}^1).
\]

Boundaries for \(\hat{\pi}_{12}, \hat{\pi}_{13}, \hat{\pi}_{21}, \hat{\pi}_{23}, \hat{\pi}_{31},\) and \(\hat{\pi}_{32}\) can be derived as well.

### A.2 Unconstrained.

Without the monotonicity constraint, none of the parameters governing the either the joint distribution or marginal distributions of \(D(0), D(1)\) and \(Y(0, D(0)), Y(1, D(1))\) are identified. Instead, we have from (3.3)

\[
\begin{align*}
\hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} &= p_{00}^0 \\
\hat{\pi}_{13} + \hat{\pi}_{14} + \hat{\pi}_{23} + \hat{\pi}_{24} &= p_{01}^0 \\
\hat{\pi}_{31} + \hat{\pi}_{32} + \hat{\pi}_{41} + \hat{\pi}_{42} &= p_{10}^0 \\
\hat{\pi}_{33} + \hat{\pi}_{34} + \hat{\pi}_{43} + \hat{\pi}_{44} &= p_{11}^0 \\
\hat{\pi}_{11} + \hat{\pi}_{14} + \hat{\pi}_{41} + \hat{\pi}_{44} &= p_{00}^1 \\
\hat{\pi}_{12} + \hat{\pi}_{13} + \hat{\pi}_{42} + \hat{\pi}_{43} &= p_{01}^1 \\
\hat{\pi}_{21} + \hat{\pi}_{24} + \hat{\pi}_{31} + \hat{\pi}_{34} &= p_{10}^1 \\
\hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} &= p_{11}^1
\end{align*}
\]

subject to \(\sum_{i=1}^{4} \sum_{j=1}^{4} \hat{\pi}_{ij} = 1.\)

A similar derivation to that developed under monotonicity using (A.3) shows

\[
\begin{align*}
\max(0, p_{00}^0 - p_{00}^1, p_{10}^1 - p_{11}^1) &\leq \pi_{12} \leq \min(p_{00}^0, p_{01}^1) \\
\max(0, p_{01}^0 - p_{01}^1, p_{10}^1 - p_{11}^1) &\leq \pi_{14} \leq \min(p_{01}^0, p_{00}^1) \\
\max(0, p_{00}^0 - p_{00}^1, p_{01}^1 - p_{10}^0) &\leq \pi_{22} \leq \min(p_{00}^0, p_{11}^1)
\end{align*}
\]
\[
\max(0, p_{00}^0 - p_{01}^1 - p_{11}^1 - p_{10}^1) \leq \pi_{24} \leq \min(p_{00}^0, p_{11}^1)
\]
\[
\max(0, p_{10}^0 - p_{10}^1 - p_{01}^0 - p_{00}^1) \leq \pi_{32} \leq \min(p_{10}^0, p_{11}^1)
\]
\[
\max(0, p_{01}^0 - p_{00}^1 - p_{11}^0 - p_{10}^1) \leq \pi_{34} \leq \min(p_{01}^0, p_{11}^1)
\]
\[
\max(0, p_{10}^0 - p_{10}^1 - p_{11}^0 - p_{00}^1) \leq \pi_{42} \leq \min(p_{10}^0, p_{01}^1)
\]
\[
\max(0, p_{11}^0 - p_{11}^1 - p_{10}^0 - p_{01}^1) \leq \pi_{44} \leq \min(p_{11}^0, p_{10}^1)
\]

and

\[
\max(0, p_{0+1}^0 - p_{0+1}^1 - p_{1+1}^0 - p_{1+1}^1) \leq \tilde{\pi}_{2+} \leq \min(p_{0+1}^0, p_{1+1}^1)
\]
\[
\max(0, p_{1+1}^0 - p_{1+1}^1, p_{0+}^0 - p_{0+}^1) \leq \tilde{\pi}_{4+} \leq \min(p_{1+1}^0, p_{0+}^1)
\]

Because all of the components in the linear combination of \(\pi_{ij}\) that make up the associative effect are in separate MLE equations, we can obtain the upper and lower bounds for the MLE of \(ae\) by replacing the parameters with the appropriate lower or upper bounds to maximize or minimize \(ae\):

\[
\left( I((\tilde{\pi}_{2l} + \tilde{\pi}_{4l}) \leq (\tilde{\pi}_{2u} + \tilde{\pi}_{4u})(\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I((\tilde{\pi}_{2l} + \tilde{\pi}_{4l}) \geq (\tilde{\pi}_{2u} + \tilde{\pi}_{4u})(\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}),
\right.
\]
\[
\left. I((\tilde{\pi}_{2l} + \tilde{\pi}_{4l}) \geq (\tilde{\pi}_{2u} + \tilde{\pi}_{4u})(\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) + I((\tilde{\pi}_{2l} + \tilde{\pi}_{4l}) \leq (\tilde{\pi}_{2u} + \tilde{\pi}_{4u})(\tilde{\pi}_{2+l} + \tilde{\pi}_{4+l}) \right)
\]

where \(\tilde{\pi}_{lij}\) and \(\tilde{\pi}_{uij}\) correspond to the lower and upper MLE limits for \(\tilde{\pi}_{ij}\), and equivalently \(\tilde{\pi}_{li+}\) and \(\tilde{\pi}_{ui+}\) correspond to the lower and upper MLE limits for \(\tilde{\pi}_{i+}\), and \(\tilde{\pi}_{l+j}\) and \(\tilde{\pi}_{u+j}\) to the lower and upper MLE limits for \(\tilde{\pi}_{+j}\).

Similar derivations provide the unconstrained MLE bounds for \(de\) and \(me\).

**B  Appendix B: Computing a Profile Likelihood for \(\pi_{ij}\)**

A profile likelihood can be computed for any \(\pi_{ij}\) using an EM algorithm. We describe the example for \(\pi_{22}\) under monotonicity. As in Section 3.1, let the complete data consist of the
number of subjects \( m_{ij} \), where \( z \) indexes the treatment assignment and \( i \) and \( j \) correspond to the indices previously defined for the counterfactual values of \( D(0), D(1) \) and \( Y(0), Y(1) \) respectively.

The complete data likelihood is given by \( \prod_{i} \prod_{j} \left( \frac{n}{m_{ij}^{0} + m_{ij}^{1}} \right)^{m_{ij}^{0} + m_{ij}^{1} \pi_{ij}^{n}} \), and thus the complete data sufficient statistics are \( m_{ij}^{0} + m_{ij}^{1}, i, j = 1, 2, 3 \). Replacing the complete data sufficient statistics with their expected values conditional on the estimated values of \( \pi_{ij} \) at the previous iteration yields a maximization step of

\[
\pi_{12}^{(t)} = \frac{n_{11} \pi_{12}^{(t-1)} + n_{01} \pi_{12}^{(t-1)}}{n_{12}^{0} + n_{12}^{1} + \pi_{12}^{(t-1)}},
\]

\[
\pi_{13}^{(t)} = \frac{n_{01} \pi_{13}^{(t-1)} + n_{01} \pi_{13}^{(t-1)}}{n_{13}^{0} + n_{13}^{1} + \pi_{13}^{(t-1)}},
\]

\[
\pi_{21}^{(t)} = \frac{n_{11} \pi_{21}^{(t-1)} + n_{11} \pi_{21}^{(t-1)}}{n_{21}^{0} + n_{21}^{1} + \pi_{21}^{(t-1)}},
\]

\[
\pi_{23}^{(t)} = \frac{n_{01} \pi_{23}^{(t-1)} + n_{11} \pi_{23}^{(t-1)}}{n_{23}^{0} + n_{23}^{1} + \pi_{23}^{(t-1)}},
\]

\[
\pi_{31}^{(t)} = \frac{n_{01} \pi_{31}^{(t-1)} + n_{11} \pi_{31}^{(t-1)}}{n_{31}^{0} + n_{31}^{1} + \pi_{31}^{(t-1)}},
\]

\[
\pi_{32}^{(t)} = \frac{n_{10} \pi_{32}^{(t-1)} + n_{11} \pi_{32}^{(t-1)}}{n_{32}^{0} + n_{32}^{1} + \pi_{32}^{(t-1)}},
\]

where \( n_{ij} \) is the observed cell count for \( Z = z, D(z) = i \), and \( Y(z, D(z)) = j \), \( \hat{\pi}_{11} = n_{01}^{1}/n_{1} \), \( \hat{\pi}_{33} = n_{33}^{0}/n_{0} \), and \( \pi_{22} \) is fixed at \( \pi_{22}^{(t)} \). We run the EM algorithm to obtain maximum likelihood estimates for the other components of \( \pi \), normalizing the estimates to sum to 1 after each step of the algorithm. Computing the observed data likelihood using (1) at \( \hat{\pi}_{ij}, i, j \neq 2 \) at a series of values of \( \pi_{22}^{(t)} \) yields the profile likelihood for \( \pi_{22} \). Profile likelihood for other components of \( \pi \) can be obtained in a similar fashion.

REFERENCES


