Biostatistics (2010), **11**, 2, *pp.* 353–372
doi:10.1093/biostatistics/kxp060
Advance Access publication on January 25, 2010

Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes

MICHAEL R. ELLIOTT*, TRIVELLORE E. RAGHUNATHAN

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA and Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, USA
mrelliot@umich.edu

YUN LI

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA

SUMMARY

Most investigations in the social and health sciences aim to understand the directional or causal relationship between a treatment or risk factor and outcome. Given the multitude of pathways through which the treatment or risk factor may affect the outcome, there is also an interest in decomposing the effect of a treatment of risk factor into "direct" and "mediated" effects. For example, child's 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 principal strata corresponding to 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 randomized treatment assignment. Because likelihood theory is not well developed for nonidentifiable parameters, we consider a Bayesian approach that 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 analyses by using several prior distributions that make weaker assumptions than monotonicity or the exclusion restriction. We consider an application that explores the mediating effects of adult poverty on the relationship between childhood poverty and risk of death.

Keywords: Direct effect; Mediated effect; Monotonicity; Mortality; Poverty.

*To whom correspondence should be addressed.

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 life span shortened through a variety of mechanisms. One of them might be that childhood poverty causes adult poverty, which in turn leads to reduced life span via poor (adult) health care, increased stress, and other factors (Backlund *and others*, 1996), 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 life span irrespective of adult poverty status (Kauhanen *and others*, 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 *and others*, 2005), where 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 to investigate mediation were outlined by Baron and Kenny (1986). They suggest fitting linear models to the data of the form

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

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

$$E(Y \mid D = d, Z = z) = \alpha_3 + \beta_3 z + \gamma d. \tag{1.1}$$

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 these conditions are met, then D is said to mediate the effect of Z on Y. Interpreting the coefficients in the linear model becomes more problematic when the outcome is binary, particularly if confounders are included in (1.1) (Mackinnon and Dwyer, 1993). In this setting, Mackinnon *and others* (2007) treat the outcome as a coarsened version of a continuous latent variable that can be modeled as in (1.1), with numerical integration or simulation procedures used to evaluate the marginal distribution of $Y|Z$ under an explicit assumption about the distribution of $D|Z$.

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. Wang and Taylor (2002), among others, define measures of the degree to which these surrogate markers can replace the true outcomes, that is, 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 postrandomization 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 (Rosenbaum, 1984). To get around this, Robins and Greenland (1992) define direct and indirect effects in terms of potential outcomes. 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 population averaged over all values of the mediator) as 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. In this framework, Albert (2008) develops a measure to quantify the fraction of the overall treatment effect that is due to a mediator.

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, 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 prerandomization "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 *and others* (2009) 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.

Joffe and Greene (2009) provide a comparison and contrast of the direct/indirect effect approach with the principal stratification approach. We focus on the latter in this article, defining direct and mediated effects in terms of intent-to-treat (ITT) effects within the principal strata of the mediators in Section 2. 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 a specific data application, namely estimating the mediating effect of adult poverty on the relationship between childhood poverty and risk of death. Section 5 summarizes our findings and suggests future extensions of these methods. Our work provides 2 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. 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. We also consider priors that do not constrain either the principal strata or the potential outcomes in any fashion whatsoever. 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 article, 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 $Y(Z, D(Z))$. (Because we do not allow the mediator to be manipulated independently of the treatment, $Y(Z, D(Z)) \equiv Y(Z)$ for all Z; we use the notation $Y(Z, D(Z))$

Table 1. *Joint distribution of counterfactular mediator and outcome*

		$Y(Z=0,D(0)),Y(Z=1,D(1))$				
		(0, 0)	(0, 1)	(1, 1)	(1, 0)	
$D(Z=0),D(Z=1)$	(0, 0)	π_{11}	π_{12}	π_{13}	π_{14}	π_{1+}
	(0, 1)	π_{21}	π_{22}	π_{23}	π_{24}	π_{2+}
	(1, 1)	π_{31}	π_{32}	π_{33}	π_{34}	π_{3+}
	(1, 0)	π_{41}	π_{42}	π_{43}	π_{44}	π_{4+}
		π_{+1}	π_{+2}	π_{+3}	π_{+4}	1

whenever we are emphasizing the mediators role in the causal pathway between Y and 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: $P(D(0)=d_0,D(1)=d_1,Y(0)=y_0,Y(1)=y_1)=\pi_{ij}$ for $i=1$ if $d_0=d_1=0$, $i=2$ 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 4 sets of values that support the distribution of $D(Z)$ form the 4 principal strata within which we will make inference about the potential outcomes $Y(Z,D(Z))$ and $Y(1-Z,D(1-Z))$: $D(0)=D(1)=0$; $D(0)=0,D(1)=1$; $D(0)=D(1)=1$; 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 ITT effect: the contrast of the potential outcome under $Z=1$ with the potential outcome under $Z=0$: $\sum_{D(Z)}E(Y(1,D(1))-Y(0,D(0)))=E(Y(1)-Y(0))=(\pi_{+2}+\pi_{+3})-(\pi_{+3}+\pi_{+4})=\pi_{+2}-\pi_{+4}$. 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"

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

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

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

If the effect of the exposure is entirely direct, that is, unmediated through D, then,

$$\begin{aligned}E(Y(1,D(1))-Y(0,D(0))|D(1)\neq D(0))&=E(Y(1,D(1))-Y(0,D(0))|D(1)=D(0))\\&=E(Y(1)-Y(0)),\end{aligned}$$

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, that is, there is no direct effect of Z on Y, then,

$$E(Y(1,D(1))-Y(0,D(0))|D(1)=D(0))=0;\ E(Y(1,D(1))-Y(0,D(0))|D(1)\neq D(0))$$
$$=\frac{E(Y(1)-Y(0))}{P(D(1)\neq D(0))}$$

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

$$\begin{aligned}\text{me} &= \frac{\text{ae} - (\pi_{+2} - \pi_{+4})}{\frac{\pi_{+2}-\pi_{+4}}{\pi_{2+}+\pi_{4+}} - (\pi_{+2} - \pi_{+4})} \\ &= \frac{\frac{\pi_{22}+\pi_{42}-(\pi_{24}+\pi_{44})}{\pi_{+2}-\pi_{+4}} - (\pi_{2+} + \pi_{4+})}{1 - (\pi_{2+} + \pi_{4+})}.\end{aligned}$$

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 "protective" when it has no impact on adult poverty.

Throughout the remainder of this article, 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 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": $D(0) \leqslant D(1)$; 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 $Y(0, D(0)) \geqslant Y(1, D(1))$ 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 *and others* (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

$$\text{ae} = E(Y(1, D(1)) - Y(0, D(0))|D(1) \neq D(0)) = \pi_{22}/\pi_{2+}$$

and the disassociative effect to

$$\text{de} = E(Y(1, D(1)) - Y(0, D(0))|D(1) = D(0)) = (\pi_{12} + \pi_{32})/(\pi_{1+} + \pi_{3+}).$$

The mediated effect measure reduces to

$$\text{me} = \frac{\pi_{22}/\pi_{2+} - \pi_{+2}}{\pi_{+2}/\pi_{2+} - \pi_{+2}} = \frac{\pi_{22}/\pi_{+2} - \pi_{2+}}{1 - \pi_{2+}}.$$

Note that ae, de, and me all have an upper bound of 1 under monotonicity since $\pi_{22} \leqslant \pi_{2+}$, $\pi_{12} + \pi_{32} \leqslant \pi_{1+} + \pi_{3+}$, and the dissasociative effect is constrained to be nonnegative.

3. INFERENCE FOR DIRECT AND MEDIATED EFFECTS

3.1 *Under the monotonicity assumption*

We observe $D(Z = z)$ and $Y(Z = z)$ 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

$$L(\pi; n) = (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})^{n^0_{00}}(\pi_{13} + \pi_{23})^{n^0_{01}}(\pi_{31} + \pi_{32})^{n^0_{10}}\pi_{33}^{n^0_{11}}$$
$$\times \pi_{11}^{n^1_{00}}(\pi_{12} + \pi_{13})^{n^1_{01}}(\pi_{21} + \pi_{31})^{n^1_{10}}(\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33})^{n^1_{11}}, \tag{3.1}$$

where n^z_{ij} 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 $p^z_{ij} = n^z_{ij}/n^z$. Unique maximum likelihood estimates (MLEs) for all marginal row and column percentages are available: $\hat{\pi}_{+1} = p^1_{+0}$, $\hat{\pi}_{+3} = p^0_{+1}$, and $\hat{\pi}_{+2} = 1 - \hat{\pi}_{+1} - \hat{\pi}_{+3}$; similarly, $\hat{\pi}_{1+} = p^1_{0+}$, $\hat{\pi}_{3+} = p^0_{1+}$, and $\hat{\pi}_{2+} = 1 - \hat{\pi}_{1+} - \hat{\pi}_{3+}$. Unique MLEs for the upper-left and lower-right cell parameters (π_{11} and π_{33}) can also be identified as p^1_{00} and p^0_{11}, 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 parameters, MLEs for the remaining 6 parameters can be identified up to boundaries (Chiba *and others*, 2007). In particular,

$$\max(0, p^0_{00} + p^1_{11} - 1) \leqslant \hat{\pi}_{22} \leqslant \min(p^0_{00} - p^1_{00} + \min(0, p^1_{10} - p^0_{10}), p^1_{11} - p^0_{11} + \min(0, p^1_{01} - p^0_{01})).$$

$$\max(0, p^0_{10} - p^1_{00}, p^1_{01} - p^0_{01}) \leqslant \hat{\pi}_{12} + \hat{\pi}_{32} \leqslant p^0_{10} + \min(p^0_{00} - (p^1_{00} + p^1_{10}), p^1_{01}).$$

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

$$\left(\frac{\max(0, p^0_{00} + p^1_{11} - 1)}{1 - p^1_{+0} - p^0_{+1}}, \frac{\min(p^0_{00} - p^1_{00} + \min(0, p^1_{10} - p^0_{10}), p^1_{11} - p^0_{11} + \min(0, p^1_{01} - p^0_{01}))}{1 - p^1_{+0} - p^0_{+1}}\right)$$

for the disassociative effect by

$$\left(\frac{\max(0, p^0_{10} - p^1_{00}, p^1_{01} - p^0_{01})}{p^1_{+0} + p^0_{+1}}, \frac{p^0_{10} + \min(p^0_{00} - (p^1_{00} + p^1_{10}), p^1_{01})}{p^1_{+0} + p^0_{+1}}\right)$$

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

		Y	
	D	0	1
$Z = 0$	0	n^0_{00} $(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})$	n^0_{01} $(\pi_{13} + \pi_{23})$
	1	n^0_{10} $(\pi_{31} + \pi_{32})$	n^0_{11} (π_{33})
$Z = 1$	0	n^1_{00} (π_{11})	n^1_{01} $(\pi_{12} + \pi_{13})$
	1	n^1_{10} $(\pi_{21} + \pi_{31})$	n^1_{11} $(\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33})$

and for the mediated effect by

$$\left(\frac{\max(0,\, p^0_{00} + p^1_{11} - 1)/(p^1_{+1} - p^0_{+1}) - (p^1_{1+} - p^0_{1+})}{1 - p^1_{+0} - p^0_{+1}},\right.$$

$$\left.\frac{\min(p^0_{00} - p^1_{00} + \min(0,\, p^1_{10} - p^0_{10}),\, p^1_{11} - p^0_{11} + \min(0,\, p^1_{01} - p^0_{01}))/(p^1_{+1} - p^0_{+1}) - (p^1_{1+} - p^0_{1+})}{1 - p^1_{+0} - p^0_{+1}}\right).$$

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 a scenario with equally likely never, concordant, and always mediators in which the effect of the treatment is entirely through the mediator: $\pi_{11} = 1/5$, $\pi_{12} = 0$, $\pi_{13} = 2/15$, $\pi_{21} = 0$, $\pi_{22} = 1/3$, $\pi_{23} = 0$, $\pi_{31} = 2/15$, $\pi_{32} = 0$, and $\pi_{33} = 1/5$. Figure 1 illustrates the profile likelihood for π_{22} for 3 samples: $n = 100$, $n = 500$, and $n = 2500$. (The profile likelihood is obtained by fixing π_{22} at a given value, maximizing the remaining values using an expectation-maximization (EM) algorithm, and computing the likelihood at the fixed π_{22} and the maximized values of the remaining π_{ij}; see Appendix B.)

Although some recent work has taken on the challenge for developing frequentist theory for situations in which likelihoods are flat (e.g. Romano and Shaikh, 2008), standard asymptotic methods for point estimation and interval construction do not apply. Thus, we turn to Bayesian methods to describe the

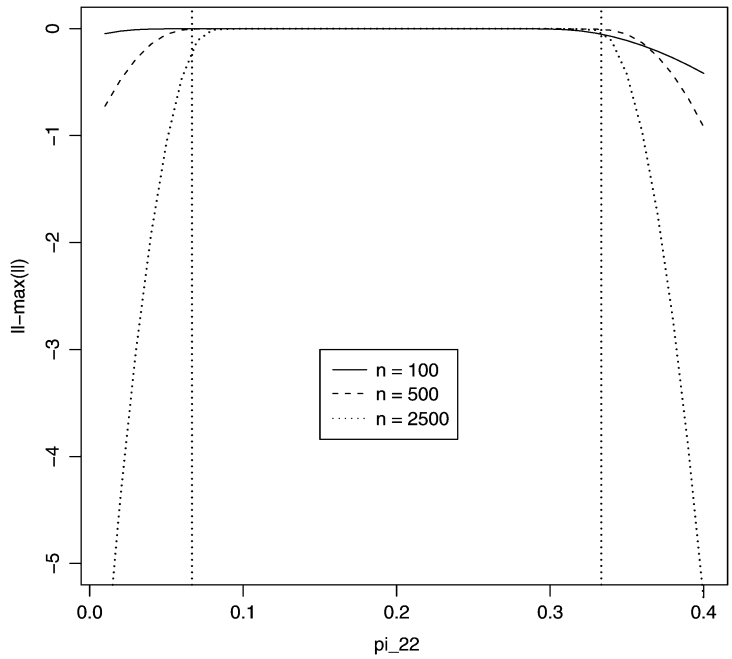

Fig. 1. Profile likelihood for $\pi_{22} = 1/3$ for 3 samples of size 100, 500, and 2500. Asymptotic boundaries for MLE given by dotted vertical lines at 1/15 and 1/3.

information available about the associative and disassociative effects of interest. We obtain simulations from the posterior distribution of $\boldsymbol{\pi}$ via a data augmentation algorithm (Tanner and Wong, 1987). Details are provided in Appendix C.

3.2 *Relaxing the monotonicity assumption*

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

$$\begin{aligned} L(\pi; n) = {} & (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})^{n^0_{00}} (\pi_{13} + \pi_{23} + \pi_{14} + \pi_{24})^{n^0_{01}} (\pi_{31} + \pi_{32} + \pi_{41} + \pi_{42})^{n^0_{10}} \\ & \times (\pi_{33} + \pi_{34} + \pi_{43} + \pi_{44})^{n^0_{11}} (\pi_{11} + \pi_{14} + \pi_{41} + \pi_{44})^{n^1_{00}} (\pi_{12} + \pi_{13} + \pi_{42} + \pi_{43})^{n^1_{01}} \\ & \times (\pi_{21} + \pi_{31} + \pi_{24} + \pi_{34})^{n^1_{10}} (\pi_{22} + \pi_{23} + \pi_{32} + \pi_{33})^{n^1_{11}}. \end{aligned} \tag{3.2}$$

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 MLEs, however. In particular, the boundary conditions for ae are

$$\left(\frac{\hat{\pi}_{22l} + \hat{\pi}_{42l} - (\hat{\pi}_{24u} + \hat{\pi}_{44u})}{I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \leqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+l} + \hat{\pi}_{4+l}) + I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \geqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+u} + \hat{\pi}_{4+u})}, \right.$$

$$\left. \frac{\hat{\pi}_{22u} + \hat{\pi}_{42u} - (\hat{\pi}_{24l} + \hat{\pi}_{44l})}{I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \geqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+l} + \hat{\pi}_{4+l}) + I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \leqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+u} + \hat{\pi}_{4+u})} \right),$$

for de are

$$\left(\frac{\hat{\pi}_{12l} + \hat{\pi}_{32l} - (\hat{\pi}_{14u} + \hat{\pi}_{34u})}{I((\hat{\pi}_{12l} + \hat{\pi}_{32l}) \leqslant (\hat{\pi}_{14u} + \hat{\pi}_{34u}))(\hat{\pi}_{1+l} + \hat{\pi}_{3+l}) + I((\hat{\pi}_{12l} + \hat{\pi}_{32l}) \geqslant (\hat{\pi}_{14u} + \hat{\pi}_{34u}))(\hat{\pi}_{1+u} + \hat{\pi}_{3+u})}, \right.$$

$$\left. \frac{\hat{\pi}_{12u} + \hat{\pi}_{32u} - (\hat{\pi}_{14l} + \hat{\pi}_{34l})}{I((\hat{\pi}_{12l} + \hat{\pi}_{32l}) \geqslant (\hat{\pi}_{14u} + \hat{\pi}_{34u}))(\hat{\pi}_{1+l} + \hat{\pi}_{3+l}) + I((\hat{\pi}_{12l} + \hat{\pi}_{32l}) \leqslant (\hat{\pi}_{14u} + \hat{\pi}_{34u}))(\hat{\pi}_{1+u} + \hat{\pi}_{3+u})} \right),$$

and for me are

$$\frac{A - B}{1 - I(A \leqslant B)(\hat{\pi}_{2+l} + \hat{\pi}_{4+l}) + I(A \geqslant B)(\hat{\pi}_{2+u} + \hat{\pi}_{4+u})},$$

where

$$A = \frac{\hat{\pi}_{22l} + \hat{\pi}_{42l} - (\hat{\pi}_{24u} + \hat{\pi}_{44u})}{I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \leqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{+2l} + \hat{\pi}_{+4l}) + I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \geqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{+2u} + \hat{\pi}_{+4u})},$$
$$B = (\hat{\pi}_{2+u} + \hat{\pi}_{4+u})$$

and the lower and upper MLE limits $\hat{\pi}_{lij}$ and $\hat{\pi}_{uij}$ for $\hat{\pi}_{ij}$, and equivalently $\hat{\pi}_{li+}$ and $\hat{\pi}_{ui+}$ for $\hat{\pi}_{i+}$, and $\hat{\pi}_{l+j}$ and $\hat{\pi}_{u+j}$ for $\hat{\pi}_{+j}$ are provided in Appendix A.

3.3 *Stochastic monotonicity assumption*

We also consider a restricted prior that constrains $\pi_{2+} \geqslant \pi_{4+}$ and $\pi_{j2} \geqslant \pi_{j4}$ for $j = 1, 2, 3$, requiring that the fraction of "concordant" mediators be greater than the fraction of "discordant" mediators, and that, within all 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 $\boldsymbol{\pi}$ 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. This prior implies an upper limit of 1 for the mediated effect me since negative disassociative effects are no longer possible.

Closed-form solutions for the MLE boundaries for the ae, de, and me can no longer be obtained; linear programming methods may be used instead (Balke and Pearl, 1997).

4. 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 6928 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 3352 respondents in 2000. Survival status by childhood poverty status and adult poverty status is shown in Table 3: 28% of children not in poverty were adults in poverty, and 44% of children in poverty were also adults in poverty.

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 martial 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 4).

Table 3. *Childhood poverty status, adult poverty status, and survival status of Alameda County Study subjects*

	Survival status		
	Alive	Dead	
Not in childhood poverty			
Not in poverty as adult	1329	254	1583
In poverty as adult	507	122	629
	1836	376	2212
In childhood poverty			
Not in poverty as adult	504	130	634
In poverty as adult	369	137	506
	873	267	1140

Table 4. *Log OR of being in childhood poverty, unadjusted and adjusted for propensity score quintile (standard error in parenthesis)*

	Unadjusted	Adjusted
Female	−0.045 (0.073)	0.010 (0.076)
Age (years)	0.036 (0.005)	−0.009 (0.009)
African-American (versus white)	0.930 (0.117)	0.552 (0.146)
Other (versus white)	−0.374 (0.191)	0.023 (0.200)

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% CI 0.93, 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 5 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 stratifying the data by propensity score quintile and running separate Markov chain Monte Carlo chains within each stratum. A draw from the posterior of $\boldsymbol{\pi}$ is obtained as the weighted average of draws from each of the 5 propensity strata, weighted in proportion to the fraction of the sample contained in each (approximate) quintile. MLE boundaries under the stochastic monotonicity assumption are obtained using the linear programming package simplex in R (R Version 2.8.0, The R Foundation for Statistical Computing).

Allowing for nonmonotonicity 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 (0.076, 95% CI −0.094, 0.237) and disassociative effects (0.054, 95% CI −0.074, 0.179) are approximately equal, with no strong evidence of mediation effects (0.16, 95% CI −1.86, 2.28), although the possibility cannot be discounted due to wide credible intervals. Constraining the prior decreases the posterior median of the associative effect (0.057, 95% CI −0.026, 0.153) and increases the posterior median of the disassociative effect (0.076, 95% CI 0.021, 0.161), 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. Balancing on age and race via the propensity score analysis moves the posterior median of the associative effect 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 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 that the fraction of never mediators is 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

Table 5. *Posterior* 5*th,* 50*th, and* 95*th 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*

	Unadjusted			Adjusted		
	5th	50th	95th	5th	50th	95th
Unconstrained (MLE)						
ae (−1.07, 1.46)	−0.094	0.076	0.237	−0.044	0.035	0.114
de (−0.62, 0.86)	−0.074	0.054	0.179	−0.030	0.034	0.098
me (−4.07, 4.20)	−1.857	0.159	2.28	−2.32	0.018	2.31
Never mediators (0.271, 0.556)	0.329	0.416	0.502	0.353	0.402	0.451
Concordant mediators (0.159, 0.444)	0.211	0.298	0.384	0.258	0.308	0.356
Always mediators (0, 0.284)	0.062	0.145	0.229	0.106	0.148	0.192
Discordant mediators (0, 0.284)	0.057	0.139	0.225	0.098	0.141	0.185
Stochastic monotonicity (MLE)						
ae (−1.07, 1.46)	−0.026	0.057	0.153	−0.008	0.035	0.082
de (−0.62, 0.86)	0.021	0.076	0.161	0.046	0.076	0.116
me (−4.07, 4.20)	−1.346	−0.117	0.672	−1.066	−0.331	0.172
Never mediators (0.271, 0.556)	0.320	0.413	0.504	0.347	0.397	0.447
Concordant mediators (0.159, 0.444)	0.209	0.302	0.393	0.262	0.313	0.362
Always mediators (0, 0.284)	0.055	0.143	0.232	0.101	0.145	0.191
Discordant mediators (0, 0.284)	0.055	0.143	0.232	0.100	0.145	0.190
Monotonicity (MLE)						
ae (0, 1)	0.010	0.125	0.330	0.055	0.117	0.201
de (0, 0.069)	0.019	0.056	0.093	0.034	0.054	0.077
me (−0.190, 1)	−0.161	0.159	0.690	−0.024	0.175	0.406
Never mediators (0.556)	0.531	0.555	0.579	0.513	0.538	0.562
Concordant mediators (0.160)	0.132	0.161	0.189	0.151	0.179	0.209
Always mediators (0.284)	0.269	0.284	0.300	0.266	0.283	0.299

(0.056, 95% CI 0.019, 0.093). The associative effect is generally centered near 0(0.125, 95% CI 0.010, 0.330), although the possibility of a substantial associative effect suggesting mediation through adult poverty cannot be entirely ruled out. Assuming monotonicity has little effect on the posterior median for the mediation effect but does shrink the posterior intervals for the mediation effects substantially. Balancing on age and race has little impact on the point estimates under monotonicity but substantially reduces the width of the credible intervals.

Figure 2 shows the posterior distributions of ae, de, and me adjusting for age, race, and gender under the 3 prior assumptions considered.

In sum, allowing for the possibility that some persons can somehow be "inoculated" against adult poverty by the experience of childhood poverty suggests that the effect of childhood poverty on risk of death is largely direct and not mediated through the experience of adult poverty caused by childhood poverty. Assuming monotonicity—that there are no persons who experience adult poverty if and only if they do not experience childhood poverty—provides very modest evidence that the increased risk of death among persons experiencing childhood poverty is partly mediated through adult poverty. Standard regression methods applied to this data yield results consistent with the results obtained under monotonicity, although the data suggest that a modest fraction of the population may indeed by inoculated.

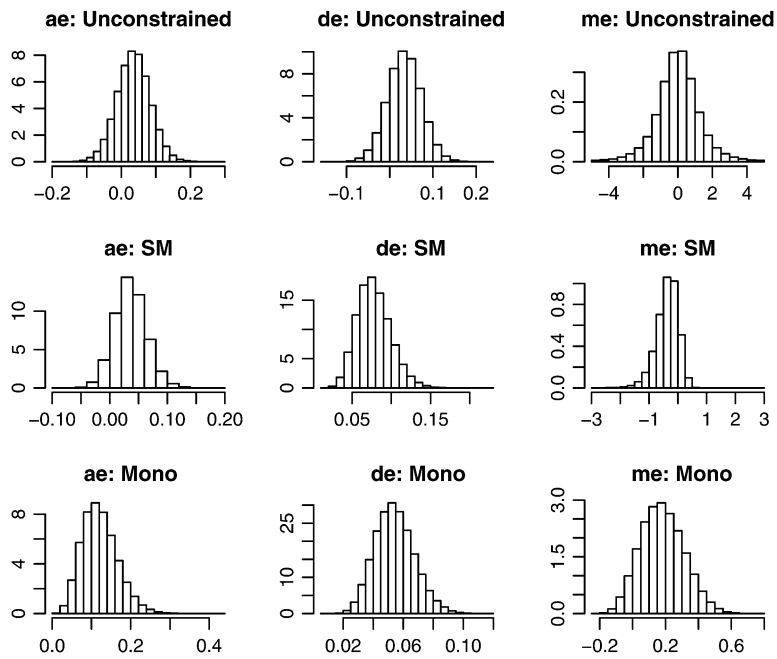

Fig. 2. Posterior distribution of associative (ae), disassociative (de), and mediated effect (me) of childhood poverty of risk of death mediated through adult poverty, adjusted for age, race, and gender using propensity scores. SM = stochastic monotonicity assumption; Mono = monotonicity assumption.

5. DISCUSSION

Standard regression approaches to mediation such as Baron and Kenny (1986) lack causal interpretation due to potential unobserved confounding even when treatment is randomized because mediator is observed postrandomization. Use of principal strata defined using the counterfactual distribution of the mediators creates a conceptual prerandomization variable. In particular, a dissassociative effect of treatment can be estimated as the 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) (see simulation studies available at http://www.sph.umich.edu/∼mrelliot/causal/med-bios2.pdf). The Bayesian approach also allow us to incorporate constraints such as monotonicity or a relaxed stochastic monotonicity.

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 affect mediating behaviors but would be independent of the joint distribution of the potential outcomes. In general, we propose using propensity scores to balance on observed covariates.

A reviewer raised the issue of whether differences between the results of a standard regression analysis of mediation effects and the results of an analysis using principal stratification to account for post-trandomization selection bias have implications for study design, above and beyond the need to obtain unconfounded treatment assignment to start with. While Hudgens and Gilbert (2009) considered power and sample designs in the surrogacy setting, they did so under conditions sufficiently restrictive to obtain consistent estimators. In our setting, with flat likelihoods, sensible sample size calculations could focus on determination of interval widths under a variety of plausible assumptions about direct and mediating mechanisms. It is less clear to us now we might make use of observed discrepancies between the principal stratum and the standard regression approaches in designing further studies; however, we may be overly pessimistic in this assessment, and look forward to others' consideration of this question.

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 noncompliance 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 prove more fruitful.

ACKNOWLEDGMENTS

The author would like to thank Jeremy Taylor, Thomas Ten Have, Dylan Small, and Marshall Joffe along with the associated editor and 2 anonymous reviewers or their helpful comments. *Conflict of Interest:* None declared.

FUNDING

National Institute of Mental Health (R01MH-078016); National Cancer Institute (R01CA-129102).

APPENDIX A: DERIVATION OF BOUNDARIES FOR ASSOCIATIVE AND DISASSOCIATIVE EFFECTS

A.1 *Under monotonicity*

Consider the observed data likelihood given by (3.1) under the assumption of monotonicity for the mediator and outcome. Unique MLEs for π_{11} and π_{33} are given by $\hat{\pi}_{11} = p^1_{00}$ and $\hat{\pi}_{33} = p^0_{11}$, with the remaining MLEs identified only up to sums:

$$\hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} = p^0_{00}, \tag{A.1}$$

$$\hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} = p^1_{11} \tag{A.2}$$

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

From (A.1), we have $p^0_{00} - p^1_{00} = \hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} - \hat{\pi}_{11} = \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} \geqslant \hat{\pi}_{22}$. From (A.2), we have $p^1_{11} - p^0_{11} = \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} - \hat{\pi}_{33} = \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} \geqslant \hat{\pi}_{22}$. Putting (A.1) and (A.2) together, we have $p^0_{00} + p^1_{11} - 1 = \hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} + \hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} - 1 = \hat{\pi}_{22} - \hat{\pi}_{13} - \hat{\pi}_{31} \leqslant \hat{\pi}_{22}$. We also have from $p^0_{10} - p^1_{10} = \hat{\pi}_{31} + \hat{\pi}_{32} - (\hat{\pi}_{21} + \hat{\pi}_{31}) = \hat{\pi}_{32} - \hat{\pi}_{21}$ that

$p^0_{00} - p^1_{00} + (p^1_{10} - p^0_{10}) = \hat{\pi}_{12} + \hat{\pi}_{22} + \hat{\pi}_{32} = \hat{\pi}_{+2} \leqslant \hat{\pi}_{22}$, and similarly $p^1_{11} - p^0_{11} + (p^1_{01} - p^0_{01}) \leqslant \hat{\pi}_{22}$. Thus,

$$\max(0,\, p^0_{00} + p^1_{11} - 1) \leqslant \hat{\pi}_{22} \leqslant \min(p^0_{00} - p^1_{00} + \min(0,\, p^1_{10} - p^0_{10}),\, p^1_{11} - p^0_{11} + \min(0,\, p^1_{01} - p^0_{01})).$$

Because $\hat{\pi}_{12} + \hat{\pi}_{32} = \hat{\pi}_{+2} - \hat{\pi}_{22}$, we have

$$\hat{\pi}_{+2} - \min(p^0_{00} - p^1_{00},\, p^1_{11} - p^0_{11}) \leqslant \hat{\pi}_{12} + \hat{\pi}_{32} \leqslant \hat{\pi}_{+2} - \max(0,\, p^0_{00} + p^1_{11} - 1).$$

Using $\hat{\pi}_{+2} = 1 - p^1_{+0} - p^0_{+1} = p^0_{10} + p^0_{00} - (p^1_{00} + p^1_{10})$, we have

$$p^0_{10} + p^0_{00} - (p^1_{00} + p^1_{10}) - \min(p^0_{00} - p^1_{00},\, p^1_{11} - p^0_{11})$$

$$\leqslant \hat{\pi}_{12} + \hat{\pi}_{32} \leqslant p^0_{10} + p^0_{00} - (p^1_{00} + p^1_{10}) - \max(0,\, p^0_{00} + p^1_{11} - 1)$$

or

$$\max(0,\, p^0_{10} - p^1_{00},\, p^1_{01} - p^0_{01}) \leqslant \hat{\pi}_{12} + \hat{\pi}_{32} \leqslant p^0_{10} + \min(p^0_{00} - (p^1_{00} + p^1_{10}),\, p^1_{01}).$$

Thus, boundaries for the ae are given by

$$\left(\frac{\max(0,\, p^0_{00} + p^1_{11} - 1)}{1 - p^1_{+0} - p^0_{+1}},\, \frac{\min(p^0_{00} - p^1_{00} + \min(0,\, p^1_{10} - p^0_{10}),\, p^1_{11} - p^0_{11} + \min(0,\, p^1_{01} - p^0_{01}))}{1 - p^1_{+0} - p^0_{+1}} \right)$$

for the de by

$$\left(\frac{\max(0,\, p^0_{10} - p^1_{00},\, p^1_{01} - p^0_{01})}{p^1_{+0} + p^0_{+1}},\, \frac{p^0_{10} + \min(p^0_{00} - (p^1_{00} + p^1_{10}),\, p^1_{01})}{p^1_{+0} + p^0_{+1}} \right)$$

and for the me by

$$\left(\frac{\max(0,\, p^0_{00} + p^1_{11} - 1)/(p^1_{+1} - p^0_{+1}) - (p^1_{1+} - p^0_{1+})}{1 - p^1_{+0} - p^0_{+1}} \right).$$

A.2 *Unconstrained*

Without the monotonicity constraint, none of the parameters governing either the joint distribution or marginal distributions of $D(0)$, $D(1)$, and $Y(0, D(0))$, $Y(1, D(1))$ are identified. Instead, we have from (3.2)

$$\begin{aligned}
\hat{\pi}_{11} + \hat{\pi}_{12} + \hat{\pi}_{21} + \hat{\pi}_{22} &= p^0_{00}, \\
\hat{\pi}_{13} + \hat{\pi}_{14} + \hat{\pi}_{23} + \hat{\pi}_{24} &= p^0_{01}, \\
\hat{\pi}_{31} + \hat{\pi}_{32} + \hat{\pi}_{41} + \hat{\pi}_{42} &= p^0_{10}, \\
\hat{\pi}_{33} + \hat{\pi}_{34} + \hat{\pi}_{43} + \hat{\pi}_{44} &= p^0_{11}, \\
\hat{\pi}_{11} + \hat{\pi}_{14} + \hat{\pi}_{41} + \hat{\pi}_{44} &= p^1_{00}, \\
\hat{\pi}_{12} + \hat{\pi}_{13} + \hat{\pi}_{42} + \hat{\pi}_{43} &= p^1_{01}, \\
\hat{\pi}_{21} + \hat{\pi}_{24} + \hat{\pi}_{31} + \hat{\pi}_{34} &= p^1_{10}, \\
\hat{\pi}_{22} + \hat{\pi}_{23} + \hat{\pi}_{32} + \hat{\pi}_{33} &= p^1_{11}
\end{aligned} \tag{A.3}$$

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

$$\max(0,\, p^0_{00} - p^1_{00} - p^1_{10} - p^1_{11}) \leqslant \pi_{12} \leqslant \min(p^0_{00},\, p^1_{01}),$$

$$\max(0,\, p^0_{01} - p^1_{01} - p^1_{10} - p^1_{11}) \leqslant \pi_{14} \leqslant \min(p^0_{01},\, p^1_{00}),$$

$$\max(0,\, p^0_{00} - p^1_{00} - p^1_{01} - p^1_{10}) \leqslant \pi_{22} \leqslant \min(p^0_{00},\, p^1_{11}),$$

$$\max(0,\, p^0_{01} - p^1_{01} - p^1_{11} - p^1_{10}) \leqslant \pi_{24} \leqslant \min(p^0_{01},\, p^1_{10}),$$

$$\max(0,\, p^0_{10} - p^1_{10} - p^1_{01} - p^1_{00}) \leqslant \pi_{32} \leqslant \min(p^0_{10},\, p^1_{11}),$$

$$\max(0,\, p^0_{11} - p^1_{00} - p^1_{01} - p^1_{10}) \leqslant \pi_{34} \leqslant \min(p^0_{11},\, p^1_{10}),$$

$$\max(0,\, p^0_{10} - p^1_{10} - p^1_{11} - p^1_{00}) \leqslant \pi_{42} \leqslant \min(p^0_{10},\, p^1_{01}),$$

$$\max(0,\, p^0_{11} - p^1_{11} - p^1_{10} - p^1_{01}) \leqslant \pi_{44} \leqslant \min(p^0_{11},\, p^1_{00})$$

and

$$\max(0,\, p^0_{0+} - p^1_{0+},\, p^1_{1+} - p^0_{1+}) \leqslant \hat{\pi}_{2+} \leqslant \min(p^0_{0+},\, p^1_{1+}),$$

$$\max(0,\, p^0_{1+} - p^1_{1+},\, p^1_{0+} - p^0_{0+}) \leqslant \hat{\pi}_{4+} \leqslant \min(p^0_{1+},\, p^1_{0+}).$$

Boundary conditions for linear combinations of the latent cell probabilities can be obtained using linear programming methods (Balke and Pearl, 1997). However, because all the components in the linear combination of π_{ij} that make up the associative effect are in MLE equations with no common elements, 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:

$$\left(\frac{\hat{\pi}_{22l} + \hat{\pi}_{42l} - (\hat{\pi}_{24u} + \hat{\pi}_{44u})}{I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \leqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+l} + \hat{\pi}_{4+l}) + I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \geqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+u} + \hat{\pi}_{4+u})}, \right.$$

$$\left. \frac{\hat{\pi}_{22u} + \hat{\pi}_{42u} - (\hat{\pi}_{24l} + \hat{\pi}_{44l})}{I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \geqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+l} + \hat{\pi}_{4+l}) + I((\hat{\pi}_{22l} + \hat{\pi}_{42l}) \leqslant (\hat{\pi}_{24u} + \hat{\pi}_{44u}))(\hat{\pi}_{2+u} + \hat{\pi}_{4+u})} \right),$$

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

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

Appendix B: computing a profile likelihood for π_{ij}

A profile likelihood can be computed for any π_{ij} using an EM algorithm. We describe the example for π_{22} under monotonicity. Let the complete data consist of the number of subjects m^z_{ij}, where z indexes the treatment assignment and i and j correspond to the indices defined for the counterfactual values of $D(0)$, $D(1)$, and $Y(0), Y(1)$ in Section 2 (see Table B.1). The complete data likelihood is given by $\prod_i \prod_j \binom{n}{m^0_{ij} + m^1_{ij}} \pi_{ij}^{m^0_{ij} + m^1_{ij}}$, and thus the complete data sufficient statistics are $m^0_{ij} + m^1_{ij}$,

Table B.1. *Complete data cell counts*

		$Y(Z=0, D(0)), Y(Z=1, D(1))$			
		$(0,0)$	$(0,1)$	$(1,1)$	$(1,0)$
$Z=0$	$(0,0)$	m^0_{11}	m^0_{12}	m^0_{13}	m^0_{14}
$D(Z=0), D(Z=1)$	$(0,1)$	m^0_{21}	m^0_{22}	m^0_{23}	m^0_{24}
	$(1,1)$	m^0_{31}	m^0_{32}	m^0_{33}	m^0_{34}
	$(1,0)$	m^0_{41}	m^0_{42}	m^0_{43}	m^0_{44}
$Z=1$	$(0,0)$	m^1_{11}	m^1_{12}	m^1_{13}	m^1_{14}
$D(Z=0), D(Z=1)$	$(0,1)$	m^1_{21}	m^1_{22}	m^1_{23}	m^1_{24}
	$(1,1)$	m^1_{31}	m^1_{32}	m^1_{33}	m^1_{34}
	$(1,0)$	m^1_{41}	m^1_{42}	m^1_{43}	m^1_{44}

$i, j = 1, 2, 3$. Replacing the complete data sufficient statistics with their expected values conditional on the estimated values of π_{ij} at the previous iteration yields a maximization step of

$$\pi^{(t)}_{12} = \frac{n^0_{11}\dfrac{\pi^{(t-1)}_{12}}{\hat{\pi}_{11}+\pi^{(t-1)}_{12}+\pi^{(t-1)}_{21}+\pi^0_{22}} + n^1_{01}\dfrac{\pi^{(t-1)}_{12}}{\pi^{(t-1)}_{12}+\pi^{(t-1)}_{13}}}{n},$$

$$\pi^{(t)}_{13} = \frac{n^0_{01}\dfrac{\pi^{(t-1)}_{13}}{\pi^{(t-1)}_{13}+\pi^{(t-1)}_{23}} + n^1_{01}\dfrac{\pi^{(t-1)}_{13}}{\pi^{(t-1)}_{12}+\pi^{(t-1)}_{13}}}{n},$$

$$\pi^{(t)}_{21} = \frac{n^0_{11}\dfrac{\pi^{(t-1)}_{21}}{\hat{\pi}_{11}+\pi^{(t-1)}_{12}+\pi^{(t-1)}_{21}+\pi^0_{22}} + n^1_{10}\dfrac{\pi^{(t-1)}_{21}}{\pi^{(t-1)}_{21}+\pi^{(t-1)}_{31}}}{n},$$

$$\pi^{(t)}_{23} = \frac{n^0_{01}\dfrac{\pi^{(t-1)}_{23}}{\pi^{(t-1)}_{13}+\pi^{(t-1)}_{23}} + n^1_{11}\dfrac{\pi^{(t-1)}_{23}}{\pi^0_{22}+\pi^{(t-1)}_{23}+\pi^{(t-1)}_{32}+\hat{\pi}_{33}}}{n},$$

$$\pi^{(t)}_{31} = \frac{n^0_{10}\dfrac{\pi^{(t-1)}_{31}}{\pi^{(t-1)}_{31}+\pi^{(t-1)}_{32}} + n^1_{10}\dfrac{\pi^{(t-1)}_{31}}{\pi^{(t-1)}_{21}+\pi^{(t-1)}_{31}}}{n},$$

$$\pi^{(t)}_{32} = \frac{n^0_{10}\dfrac{\pi^{(t-1)}_{32}}{\pi^{(t-1)}_{31}+\pi^{(t-1)}_{32}} + n^1_{11}\dfrac{\pi^{(t-1)}_{32}}{\pi^0_{22}+\pi^{(t-1)}_{23}+\pi^{(t-1)}_{32}+\hat{\pi}_{33}}}{n},$$

where n^z_{ij} is the observed cell count for $Z = z$, $D(z) = i$, and $Y(z, D(z)) = j$, $\hat{\pi}_{11} = n^1_{00}/n_1$, $\hat{\pi}_{33} = n^0_{33}/n_0$, and π_{22} is fixed at π^0_{22} . We run the EM algorithm to obtain MLEs for the other components of π, normalizing the estimates to sum to 1 after each step of the algorithm. Computing the observed data likelihood using (1.1) at $\hat{\pi}_{ij}$, $i, j \neq 2$ at a series of values of π^0_{22} yields the profile likelihood for π_{22}. Profile likelihood for other components of π can be obtained in a similar fashion.

APPENDIX C: BAYESIAN INFERENCE FOR DIRECT AND MEDIATED EFFECTS

C.1 *Under monotonicity*

The complete data is given by the cell counts m_{ij}^z, where z indexes the treatment assignment and i and j correspond to the indices defined for the counterfactual values of $D(0)$, $D(1)$ and $Y(0)$, $Y(1)$ in Section 2. Under randomization and SUTVA, we have $\boldsymbol{m}^z \sim \text{MULTI}(n^z; \pi_{11}, \ldots, \pi_{33})$, so the data augmentation step is given by draws from the multinomial distribution:

$$
\begin{aligned}
m_{11}^0, m_{12}^0, m_{21}^0, m_{22}^0 &\sim \text{MULTI}\left(n_{00}^0, \frac{\pi_{11}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}, \frac{\pi_{12}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})},\right.\\
&\qquad \left.\frac{\pi_{21}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}, \frac{\pi_{22}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}\right),\\
m_{13}^0, m_{23}^0 &\sim \text{MULTI}\left(n_{01}^0, \frac{\pi_{13}}{(\pi_{13}+\pi_{23})}, \frac{\pi_{23}}{(\pi_{13}+\pi_{23})}\right),\\
m_{31}^0, m_{32}^0 &\sim \text{MULTI}\left(n_{10}^0, \frac{\pi_{31}}{(\pi_{31}+\pi_{32})}, \frac{\pi_{32}}{(\pi_{31}+\pi_{32})}\right),\\
m_{12}^1, m_{13}^1 &\sim \text{MULTI}\left(n_{01}^1, \frac{\pi_{12}}{(\pi_{12}+\pi_{13})}, \frac{\pi_{13}}{(\pi_{12}+\pi_{13})}\right),\\
m_{21}^1, m_{31}^1 &\sim \text{MULTI}\left(n_{10}^1, \frac{\pi_{21}}{(\pi_{21}+\pi_{31})}, \frac{\pi_{31}}{(\pi_{31}+\pi_{31})}\right),\\
m_{22}^1, m_{23}^1, m_{32}^1, m_{33}^1 &\sim \text{MULTI}\left(n_{11}^1, \frac{\pi_{22}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}, \frac{\pi_{23}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})},\right.\\
&\qquad \left.\frac{\pi_{32}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}, \frac{\pi_{33}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}\right).
\end{aligned}
$$

We also have $m_{33}^0 = n_{11}^0$ 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 be 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 $\boldsymbol{\pi}$ conditional on the previous draw of $\mathbf{m}$ from $\text{DIRICHLET}(m_{11}^0 + m_{11}^1 + 1, \ldots, m_{33}^0 + m_{33}^1 + 1)$.

C.2 *Relaxing the monotonicity assumption*

Here, the data augmentation step is given by

$$m^0_{11}, m^0_{12}, m^0_{21}, m^0_{22} \sim \text{MULTI}\left(n^0_{00}, \frac{\pi_{11}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}, \frac{\pi_{12}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}, \frac{\pi_{21}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}, \frac{\pi_{22}}{(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})}\right),$$

$$m^0_{13}, m^0_{23}, m^0_{14}, m^0_{24} \sim \text{MULTI}\left(n^0_{01}, \frac{\pi_{13}}{(\pi_{13}+\pi_{23}+\pi_{14}+\pi_{24})}, \frac{\pi_{23}}{(\pi_{13}+\pi_{23}+\pi_{14}+\pi_{24})}, \frac{\pi_{14}}{(\pi_{13}+\pi_{23}+\pi_{14}+\pi_{24})}, \frac{\pi_{24}}{(\pi_{13}+\pi_{23}+\pi_{14}+\pi_{24})}\right),$$

$$m^0_{31}, m^0_{32}, m^0_{41}, m^0_{42} \sim \text{MULTI}\left(n^0_{10}, \frac{\pi_{31}}{(\pi_{31}+\pi_{32}+\pi_{41}+\pi_{42})}, \frac{\pi_{32}}{(\pi_{31}+\pi_{32}+\pi_{41}+\pi_{42})}, \frac{\pi_{41}}{(\pi_{31}+\pi_{32}+\pi_{41}+\pi_{42})}, \frac{\pi_{42}}{(\pi_{31}+\pi_{32}+\pi_{41}+\pi_{42})}\right),$$

$$m^0_{33}, m^0_{34}, m^0_{43}, m^0_{44} \sim \text{MULTI}\left(n^0_{11}, \frac{\pi_{33}}{(\pi_{33}+\pi_{34}+\pi_{43}+\pi_{44})}, \frac{\pi_{34}}{(\pi_{33}+\pi_{34}+\pi_{43}+\pi_{44})}, \frac{\pi_{43}}{(\pi_{33}+\pi_{34}+\pi_{43}+\pi_{44})}, \frac{\pi_{44}}{(\pi_{33}+\pi_{34}+\pi_{43}+\pi_{44})}\right),$$

$$m^1_{11}, m^1_{14}, m^1_{41}, m^1_{44} \sim \text{MULTI}\left(n^1_{00}, \frac{\pi_{11}}{(\pi_{11}+\pi_{14}+\pi_{41}+\pi_{44})}, \frac{\pi_{14}}{(\pi_{11}+\pi_{14}+\pi_{41}+\pi_{44})}, \frac{\pi_{41}}{(\pi_{11}+\pi_{14}+\pi_{41}+\pi_{44})}, \frac{\pi_{44}}{(\pi_{11}+\pi_{14}+\pi_{41}+\pi_{44})}\right),$$

$$m^1_{12}, m^1_{13}, m^1_{42}, m^1_{43} \sim \text{MULTI}\left(n^1_{01}, \frac{\pi_{12}}{(\pi_{12}+\pi_{13}+\pi_{42}+\pi_{43})}, \frac{\pi_{13}}{(\pi_{12}+\pi_{13}+\pi_{42}+\pi_{43})}, \frac{\pi_{42}}{(\pi_{12}+\pi_{13}+\pi_{42}+\pi_{43})}, \frac{\pi_{43}}{(\pi_{12}+\pi_{13}+\pi_{42}+\pi_{43})}\right),$$

$$m^1_{21}, m^1_{31}, m^1_{24}, m^1_{34} \sim \text{MULTI}\left(n^1_{10}, \frac{\pi_{21}}{(\pi_{21}+\pi_{31}+\pi_{24}+\pi_{34})}, \frac{\pi_{31}}{(\pi_{21}+\pi_{31}+\pi_{24}+\pi_{34})}, \frac{\pi_{24}}{(\pi_{21}+\pi_{31}+\pi_{24}+\pi_{34})}, \frac{\pi_{34}}{(\pi_{21}+\pi_{31}+\pi_{24}+\pi_{34})}\right),$$

$$m^1_{22}, m^1_{23}, m^1_{32}, m^1_{33} \sim \text{MULTI}\left(n^1_{11}, \frac{\pi_{22}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}, \frac{\pi_{23}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}, \frac{\pi_{32}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}, \frac{\pi_{33}}{(\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33})}\right),$$

We retain $p(\pi) \sim \text{DIRICHLET}(1, \ldots, 1)$, so that $\pi|m \sim \text{DIRICHLET}(m^0_{11} + m^1_{11} + 1, \ldots, m^0_{44} + m^1_{44} + 1)$.

REFERENCES

ALBERT, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine* **27**, 1282–1304.

BACKLUND, E., SORLIE, P. D. AND JOHNSON, N. J. (1996). The shape of the relationship between income and mortality in the United States: evidence from the national longitudinal mortality study. *Annals of Epidemiology* **6**, 12–20.

BALKE, A. AND PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171–1176.

BARON, R. M. AND KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**, 1173–1182.

BRESLOW, L. AND KAPLAN, G. A. (1965). *Health and Ways of Living Study, 1965 Panel*: [*Alamenda county, California*]. Berkley, CA: California Department of Health Services, Human Population Laboratory.

CHIBA, Y., SATO, T. AND GREENLAND, S. (2007). Bounds on potential risks and causal risk differences under assumptions about confounding parameters. *Statistics in Medicine* **26**, 5125–5135.

FRANGAKIS, C. AND RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

GALLOP, R., SMALL, D. S., LIN, J. Y., ELLIOTT, M. R., JOFFE, M. M. AND TEN HAVE, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine* **28**, 1108–1130.

JOFFE, M. M. AND GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**, 530–538.

HABER, M. (1999). Estimation of the direct and indirect effects of vaccination. *Statistics in Medicine* **18**, 2101–2109.

HALLORAN, M. E. AND STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6**, 142–151.

HIRANO, K., IMBENS, G. W., RUBIN, D. B. AND ZHOU, X.-H (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

HUDGENS, M. G. AND GILBERT, P. B. (2009). Assessing vaccine effects in repeated low-dose challenge experiments. *Biometrics* **65**, 1223–1232.

KASS, R. E. (1989). The geometry of asymptotic inference (with discussion). *Statistical Science* **4**, 188–234.

KAUHANEN, L., LAKKA, H.-M, LYNCH, J. W. AND KAUHANEN, J. (2006). Social disadvantages in childhood and risk of all-cause death and cardiovascular disease in later life: a comparison of historical and retrospective childhood information. *International Journal of Epidemiology* **35**, 962–968.

MACKINNON, D. P. AND DWYER, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review* **17**, 144–158.

MACKINNON, D. P., LOCKWOOD, C. M., BROWN, C. H., WANG, W. AND HOFFMAN, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* **4**, 449–513.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

ROBINS, J. M. AND GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155.

ROMANO, J. P. AND SHAIKH, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference* **138**, 2786–2807.

ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147**, 656–666.

ROSENBAUM, P. R. AND RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* **6**, 34–58.

RUBIN, D. B. (1990). Comment on Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.

RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161–170.

TANNER, M. A. AND WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

TAYLOR, J. M. G., WANG, Y. AND THIEBAUT, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**, 1102–1111.

WANG, Y. AND TAYLOR, J. M. G. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58**, 803–812.

[*Received June 29, 2009; revised December 14, 2009; accepted for publication December 15, 2009*]