BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

Transparent parametrizations of models for potential outcomes

THOMAS S. RICHARDSON, ROBIN J. EVANS
University of Washington
{thomasr, rje42}@uw.edu
JAMES M. ROBINS
Harvard School of Public Health
robins@hsph.harvard.edu

SUMMARY

We consider causal models involving three binary variables: a randomized assignment Z, an exposure measure X, and a final response Y. We focus particular attention on the situation in which there may be confounding of X and Y, while at the same time measures of the effect of X on Y are of primary interest. In the case where Z has no effect on Y, other than through Z this is the instrumental variable model. Many causal quantities of interest are only partially identified. We first show via examples that the resulting posteriors may be highly sensitive to the specification of the prior distribution over compliance types. We present a 'transparent' re-parametrization of the likelihood that addresses this problem by separating the identified and non-identified parts of the parameter.

1. INTRODUCTION

The potential outcomes model for causal inference is a well-established framework for formalizing causal assumptions and modelling causal effects. However, in many contexts the causal estimands of interest are not identified by the observed data. Even in the asymptotic limit, there may be a range of values for the parameter(s) of interest that are logically possible, such parameters are often referred to as being *partially identified*.

It has been proposed by several authors to apply a standard Bayesian prior to posterior analysis to such models. It is often argued that the issue of identifiability is of secondary importance in a Bayesian analysis, provided that the posterior is 'informed' by the observed data.

Following Leamer (1978), Gustafson (2005) and Greenland (2005) we argue that partially identified models should be re-parameterized so as to separate the (wholly) identified and (wholly) non-identified parameters. Such an approach facilitates 'transparency', allowing a reader to see clearly which parts of the analysis have been informed by the data. In addition it makes it simpler for someone to incorporate their own prior beliefs that may differ from those of the analyst.

In this paper we first motivate the approach by considering a simple instrumental variable model for a randomized trial with non-compliance. We then extend this approach to the analysis of a randomized encouragement design, though still with binary treatment and response, under a variety of assumptions. Finally we develop smooth parametrizations that permit this approach to be applied in the context of continuous or discrete baseline covariates.

The paper is organized as follows: in Section 2 we introduce the notation and the basic potential outcomes model that we consider throughout. In Section 3 we motivate our approach via a simple example, and show how the method applies. In Section 4 we describe eight causal models and explicitly characterize each of them. In Section 5 we extend the approach to incorporate baseline covariates.

2. BASIC CONCEPTS

Throughout this paper we consider potential outcomes models involving three binary variables, X, Y and Z. Where:

Z is a treatment, presumed to be randomized e.g. the assigned treatment;

X is an exposure subsequent to treatment assignment;

Y is the response.

For Z we will use 1 to indicate assignment to drug, and 0 otherwise. For X we use 1 to indicate that the drug is received and 0 if not. For Y we take 1 to indicate a desirable outcome, such as survival.

The potential outcome X_z is the treatment a patient would receive if assigned to $Z = z$. We follow convention by referring to the four *compliance* types as shown in Table 1. We will use $\mathfrak{t}_X$ to denote a generic compliance type, and $\mathbb{D}_X$ the set of such types.

Table 1: *Compliance types describing the potential outcomes X_z*

$X_{z=0}$	$X_{z=1}$	Compliance Type	
0	0	Never Taker	NT
0	1	Complier	CO
1	0	Defier	DE
1	1	Always Taker	AT

Similarly we consider the four potential outcomes Y_{xz} with $x, z \in \{0, 1\}$ for Y. These describe the outcome for a given patient if they were to be assigned to $Z = z$ and then were exposed to $X = x$. For a given individual we will refer to the 4-vector of values taken by the variables $(Y_{00}, Y_{01}, Y_{10}, Y_{11})$ as their *response type*, $\mathfrak{t}_Y$. We use $\mathbb{D}_Y$ to indicate the set of such types, of which there are $2^4 = 16$ in general, though we will often consider models in which some of these are assumed to be identical.

Since we suppose the potential outcomes are well-defined, if $Z = z$ then $X = X_z$, similarly if $X = x$ and $Z = z$ then $Y = Y_{xz}$. This is referred to as the 'consistency assumption' (or axiom).

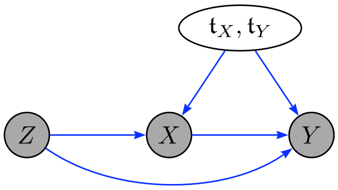


Figure 1: *Graphical representation of the model given by assumption* (1)*. The shaded nodes are observed. In this model* $\mathfrak{t}_X$ *takes* 4 *states, while* $\mathfrak{t}_Y$ *takes* 16*.*

Notation

Let $\pi_{\mathfrak{t}_X} \equiv p(\mathfrak{t}_X)$ denote the marginal probability of a given compliance type $\mathfrak{t}_X \in \mathbb{D}_X$, and

$$\pi_X \equiv \{\pi_{\mathfrak{t}_X} \mid \mathfrak{t}_X \in \mathbb{D}_X\}$$

denote a distribution on $\mathbb{D}_X$. Similarly we use $\pi_{\mathfrak{t}_Y|\mathfrak{t}_X} \equiv p(\mathfrak{t}_Y \mid \mathfrak{t}_X)$ to denote the probability of a given response type within the sub-population of individuals of compliance type $\mathfrak{t}_X$, and $\pi_{Y|X}$ to indicate a specification of all these conditional probabilities:

$$\pi_{Y|X} \equiv \{\pi_{\mathfrak{t}_Y|\mathfrak{t}_X} \mid \mathfrak{t}_X \in \mathbb{D}_X, \mathfrak{t}_Y \in \mathbb{D}_Y\}.$$

We will use π to indicate a joint distribution $p(\mathfrak{t}_X, \mathfrak{t}_Y)$ on $\mathbb{D}_X \times \mathbb{D}_Y$.

We use $\gamma^{ij}_{\mathfrak{t}_X}$ for the probability of recovery for a patient of a given compliance type $\mathfrak{t}_X$, under an intervention that sets $X = i$ and $Z = j$:

$$\gamma^{ij}_{\mathfrak{t}_X} \equiv p(Y_{x=i,z=j} = 1 \mid \mathfrak{t}_X), \text{ for } i, j \in \{0, 1\} \text{ and } \mathfrak{t}_X \in \mathbb{D}_X.$$

In places we will make use of the following compact notation for probability distributions:

$$\begin{aligned} p(y_k|x_j z_i) &\equiv p(Y = k \mid X = j, Z = i), \\ p(x_j|z_i) &\equiv p(X = j \mid Z = i), \\ p(y_k, x_j|z_i) &\equiv p(Y = k, X = j \mid Z = i). \end{aligned}$$

Finally we use Δ_k to indicate the simplex of dimension k.

Randomization assumption

We will make the randomization assumption that the distribution of types $\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle$ is the same in both the $Z = 0$ and $Z = 1$ arms:

$$Z \perp\!\!\!\perp \{X_{z=0}, X_{z=1}, Y_{x=0,z=0}, Y_{x=1,z=0}, Y_{x=1,z=0}, Y_{x=1,z=1}\}. \tag{1}$$

A causal graph corresponding to the model given by (1) is shown in Figure 1.

3. A SIMPLE MOTIVATING EXAMPLE

Pearl (2000) and Chickering and Pearl (1996) use potential outcomes to analyze the data in Table 3 which arises from a randomized trial of Cholestyramine; see Efron

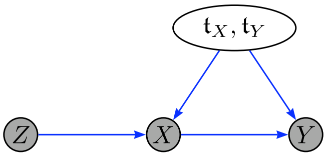


Figure 2: *Graphical representation of the IV model given by assumptions* (2) *and* (1)*. In this model* $\mathfrak{t}_X$ *takes* 4 *states, while* $\mathfrak{t}_Y$ *takes* 4*.*

and Feldman (1991). Compliance was originally measured as a percentage of prescribed dosage consumed; this measure was then dichotomized by Pearl. Similarly the response was also dichotomized to indicate a reduction in cholesterol of at least 28 units.

The potential outcomes analysis here is simplified since subjects in the control arm had no access to treatment. Hence $Z = 0$ implies $X = 0$ so there are only two compliance types (NT, CO). Since it is a randomized trial Pearl also assumes that Z has no effect on Y other than through X, or more formally:

$$Y_{xz} = Y_{xz'} \qquad \text{for all } x, z, z' \in \{0, 1\}. \tag{2}$$

In this case there are only four response types $\mathfrak{t}_Y$; see Table 2. Consequently there are eight combinations for $(\mathfrak{t}_X, \mathfrak{t}_Y) \in \{\text{NT}, \text{CO}\} \times \{HE, HU, AR, NR\}$.

When equation (2) holds we will use $Y_{x\cdot}$ to refer to $Y_{x,z=1} = Y_{x,z=0}$. Similarly we let $\gamma^{i\cdot}_{\mathfrak{t}_X} \equiv P(Y_{x=i\cdot} = 1 \mid \mathfrak{t}_X)$.

Table 2: *Response types under the exclusion restriction (2).*

$Y_{x=0\cdot}$	$Y_{x=1\cdot}$	Response Type	
0	0	Never Recover	*NR*
0	1	Helped	*HE*
1	0	Hurt	*HU*
1	1	Always Recover	*AR*

Pearl (2000) takes as his primary quantity of interest the (global) average causal effect of X on Y:

$$\text{ACE}(X \to Y) \equiv E[Y_{x=1\cdot} - Y_{x=0\cdot}] = \pi(HE) - \pi(HU).$$

Pearl proposes analyzing the model by placing a prior distribution over $p(\mathfrak{t}_X, \mathfrak{t}_Y)$ and then using Gibbs sampling to sample from the resulting posterior distribution for $\text{ACE}(X \to Y)$. He notes that the resulting posterior appears sensitive to the prior distribution and suggests that a sensitivity analysis be used.

Figure 3 illustrates this sensitivity. The solid green and red lines in the left plot show, respectively, the prior and posterior for $\text{ACE}(X \to Y)$ under a uniform $\text{Dir}(1, \ldots, 1)$ on the distribution $\pi(\mathfrak{t}_X, \mathfrak{t}_Y)$; the dashed green and red lines indicate the corresponding prior and posterior after increasing the parameter corresponding to (NT,*HE*) to 1.2, while reducing that for (NT,*NR*) to 0.8, but leaving all others at 1. If the model were identified we would expect such a change in the prior to

Table 3: *Lipid / Cholestyramine Data; originally considered by Efron and Feldman (1991); dichotomized by Pearl. There are two structural zeros.*

z	x	y	count	z	x	y	count
0	0	0	158	1	0	0	52
0	0	1	14	1	0	1	12
0	1	0	0	1	1	0	23
0	1	1	0	1	1	1	78
			172				165

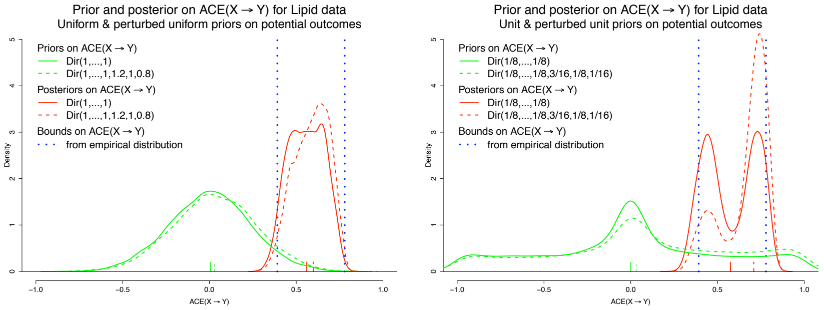


Figure 3: *Prior to posterior analysis for* ACE($X \rightarrow Y$) *for the Lipid Data; priors are green; posteriors are red; vertical lines indicate bounds on the ACE evaluated at the empirical distribution. Tick marks indicate respective medians. See text for further details.*

have little effect (the smallest observed count is 12). However, as the plot shows, this perturbation makes a considerable difference to the posterior.

Experts whom we consulted, noting the fact that there was relatively little prior support in the range dominated by the posterior, hypothesized that the sensitivity might be due to an insufficiently diffuse prior. It was suggested that a 'unit information' prior should be used instead. The right plot in Figure 3 shows the prior and posterior for the ACE resulting from a Dir$(1/8, \ldots, 1/8)$ and under a prior in which the parameter for (NT,*HE*) is increased to 3/16 while that for (NT,*NR*) is reduced to 1/16. The plot shows that the more diffuse prior on $\pi(\mathfrak{t}_X, \mathfrak{t}_Y)$ has succeeded in increasing the spread of the prior for ACE($X \rightarrow Y$), but this has come at the expense of multi-modality in the posterior, and greater prior sensitivity: notice the difference between the posterior medians (indicated at the base of the plot).

On closer inspection the sensitivity should not be surprising, since the observed data contain no information allowing us to learn about the ratio of (NT,*HE*) to (NT,*NR*): patients who are of type 'Helped' (*HE*), and 'Never Recover' (*NR*) will both have $Y_{x=0} = 0$; they only differ with respect to their values of $Y_{x=1}$. However, patients who are 'Never Takers' will never expose themselves to treatment, so these potential outcomes are never observed (at least not without instituting a new experimental protocol that eliminates non-compliance). Of course, the proportion of

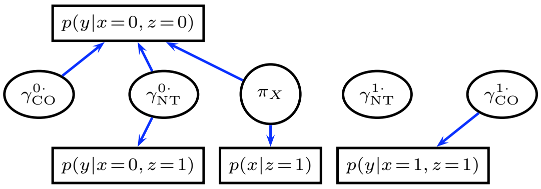

$p(y|x=0,z=0)$
$\gamma^{0\cdot}_{\mathrm{CO}}$
$\gamma^{0\cdot}_{\mathrm{NT}}$
π_X
$\gamma^{1\cdot}_{\mathrm{NT}}$
$\gamma^{1\cdot}_{\mathrm{CO}}$
$p(y|x=0,z=1)$
$p(x|z=1)$
$p(y|x=1,z=1)$

Figure 4: *A graph representing the functional dependencies in the analysis of the simple IV model with no Always Takers or Defiers. Rectangular nodes are observed; oval nodes are unknown parameters.* $p(x=1|z=0)=0$, *so* $p(y|x=1,z=0)$ *is undefined, hence these nodes are omitted.*

patients who are of type 'Helped' (rather than 'Never Recover') is directly relevant to ACE($X \to Y$).

Separating the identified from the unidentified

Figure 4 provides a graphical depiction of the functional relations between the parameters π_X, $\gamma^{i\cdot}_{\mathrm{CO}}$, and $\gamma^{i\cdot}_{\mathrm{NT}}$, and the observed distribution $p(y,x|z)$. The parameters π_X, and $\gamma^{1\cdot}_{\mathrm{CO}}$, $\gamma^{0\cdot}_{\mathrm{CO}}$, and $\gamma^{0\cdot}_{\mathrm{NT}}$ are identified thus:

$$\begin{aligned}
&\pi_{\mathrm{CO}} = p_{x_1|z_1}, \quad \gamma^{1\cdot}_{\mathrm{CO}} = p_{y_1|x_1,z_1}, \quad \gamma^{0\cdot}_{\mathrm{CO}} = (p_{y_1,x_0|z_0} - p_{y_1,x_0|z_1})/p_{x_1|z_1},\\
&\pi_{\mathrm{NT}} = p_{x_0|z_1}, \qquad\qquad\qquad\qquad\quad\; \gamma^{0\cdot}_{\mathrm{NT}} = p_{y_1|x_0,z_1}.
\end{aligned}$$

The equation for $\gamma^{0\cdot}_{\mathrm{CO}}$ leads to the following restrictions on the distribution $p(y,x|z)$:

$$\begin{aligned}
\gamma^{0\cdot}_{\mathrm{CO}} \le 1 \quad &\Rightarrow \quad p_{y_0,x_0|z_1} \le p_{y_0,x_0|z_0},\\
\gamma^{0\cdot}_{\mathrm{CO}} \ge 0 \quad &\Rightarrow \quad p_{y_1,x_0|z_1} \le p_{y_1,x_0|z_0}.
\end{aligned} \tag{3}$$

It is not hard to show that these inequalities define the set of distributions $p(y,x|z)$ arising from this potential outcome model. Consequently we may parametrize the identifiable portion of the model directly via the set of distributions $p(y,x|z)$ that obey the inequalities on the right of (3). Under a Dirichlet prior over the observed distribution $p(y,x|z)$, truncated so as to remove distributions violating (3), the posterior may easily be sampled from via conjugacy and Monte-Carlo rejection sampling.

As a by-product we may also examine the posterior probability assigned to the model defining restrictions (3) being violated under a uniform prior on the saturated model. For the Lipid data, under this prior, the posterior probability of such a violation is still 0.38. (The prior probability of violating (3) is 0.5.) This might cast doubt on the exclusion restrictions, Eq. (2). One possible explanation for a violation of Eq. (2), even in the context of a double blind study, is the dichotomization of the compliance measure; see Robins et al. (2009); Balke and Pearl (1997). Note that if the posterior probability of (3) holding is high this does not imply that the posterior

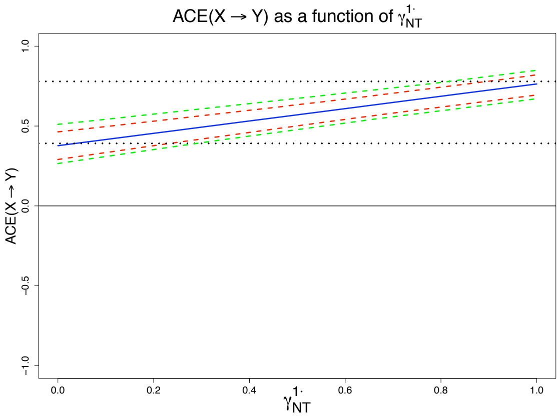


Figure 5: *The posterior for the* $\mathrm{ACE}(X \rightarrow Y)$ *for the Lipid data displayed as a function of the (completely) unidentified parameter* $\gamma_{\mathrm{NT}}^{1\cdot}$*:* (blue) *posterior median;* (red) 2.5% *and* 97.5% *quantiles;* (green) *simultaneous* 95% *posterior region obtained from a* 95% *HPD region for* $p(y, x|z)$*; horizontal lines are bounds on the ACE evaluated at the empirical distribution. A uniform prior was used on distributions* $p(y, x \mid z)$ *that satisfy the inequalities* (3).

probability of (2) is high, since the model in which (2) is violated is of the same dimension, and contains that in which it holds.

In this example we could have used $(\pi_X, \gamma_{\mathrm{CO}}^{1\cdot}, \gamma_{\mathrm{CO}}^{0\cdot}, \gamma_{\mathrm{NT}}^{0\cdot})$ rather than $p(y, x|z)$ to parametrize the identifiable part of the model. However, this approach does not generalize to more complex potential outcome models such as those that include Defiers, or make fewer exclusion restrictions, since both π_X and $\gamma_{\mathfrak{t}_X}^{i\cdot}$ may themselves be partially identified; see Richardson and Robins (2010).

Posterior distributions for the ACE

The $\mathrm{ACE}(X \rightarrow Y)$ depends on the (wholly) unidentified parameter $\gamma_{\mathrm{NT}}^{1\cdot}$:

$$\mathrm{ACE}(X \rightarrow Y) = \pi_{\mathrm{CO}}(\gamma_{\mathrm{CO}}^{1\cdot} - \gamma_{\mathrm{CO}}^{0\cdot}) + \pi_{\mathrm{NT}}(\gamma_{\mathrm{NT}}^{1\cdot} - \gamma_{\mathrm{NT}}^{0\cdot}).$$

We elect to display the posterior for $\mathrm{ACE}(X \rightarrow Y)$ as a function of $\gamma_{\mathrm{NT}}^{1\cdot}$; see Figure 5. This permits readers to see clearly the dependence of the ACE on this parameter, and to incorporate easily their priors regarding $\gamma_{\mathrm{NT}}^{1\cdot}$.

4. THE GENERAL FRAMEWORK

We now consider the general setting in which we do not assume Eq. (2), nor do we rule out the possibility of Always Takers or Defiers. Thus there are 4 × 16 possible values for $(\mathfrak{t}_X, \mathfrak{t}_Y)$.

Following Hirano et al. (2000) we consider models under which (1) holds, and (combinations of) the following three assumptions hold:

(Mon$_X$) *Monotonicity of Compliance*: $X_0 \leq X_1$, or equivalently, there are no Defiers.

(Ex$_{\mathrm{NT}}$) *Stochastic Exclusion for* NT *under non-exposure*: $\gamma^{01}_{\mathrm{NT}} = \gamma^{00}_{\mathrm{NT}}$, so among Never Takers the distributions of Y_{00} and Y_{01} are the same.

(Ex$_{\mathrm{AT}}$) *Stochastic Exclusion for* AT *under exposure*: $\gamma^{11}_{\mathrm{AT}} = \gamma^{10}_{\mathrm{AT}}$, so among Always Takers the distributions of Y_{10} and Y_{11} are the same.

Note that assumption (2) implies stochastic exclusion for all compliance types under all exposures, i.e. $\gamma^{ij}_{\mathfrak{t}_X} = \gamma^{ij'}_{\mathfrak{t}_X}$ for all $i, j, j' \in \{0, 1\}$ and all $\mathfrak{t}_X \in \mathbb{D}_X$. Figure 8 and Table 4 list these eight models. Imposing other exclusion restrictions, besides Ex$_{\mathrm{AT}}$ or Ex$_{\mathrm{NT}}$, will correspond to merely relabelling a single node $\gamma^{ij}_{\mathfrak{t}_X}$ in Figure 8 with $\gamma^{i\cdot}_{\mathfrak{t}_X}$. Thus, although the causal interpretation of estimands may change, the implied set of compatible distributions $p(y, x|z)$ will not.

The saturated model $p(y, x|z)$ consists of the cartesian product of two three dimensional simplices: $\Delta_3 \times \Delta_3$. The other seven models are all characterized by simple inequality restrictions on this set.

Inequalities defining models with Defiers

Results of Bonet (2001) imply that the set of distributions arising from a potential outcomes model satisfying (1), Ex$_{\mathrm{AT}}$ and Ex$_{\mathrm{NT}}$ may be characterized via the following inequalities:

$$p(y_0, x_0 \mid z_0) + p(y_1, x_0 \mid z_1) \leq 1, \quad p(y_1, x_0 \mid z_0) + p(y_0, x_0 \mid z_1) \leq 1, \tag{4}$$

$$p(y_0, x_1 \mid z_0) + p(y_1, x_1 \mid z_1) \leq 1, \quad p(y_0, x_1 \mid z_1) + p(y_1, x_1 \mid z_0) \leq 1. \tag{5}$$

Note that any distribution $p(y, x \mid z)$ can violate at most one of these four inequalities. In addition they are invariant under relabelling of any variable. Cai et al. (2008) give a simple interpretation of the inequalities in terms of bounds on average controlled direct effects in the potential outcomes model that only assumes (1):

$$p(y_0, x_i \mid z_0) + p(y_1, x_i \mid z_1) - 1 \leq \mathrm{ACDE}(x_i) \leq 1 - p(y_0, x_i \mid z_1) - p(y_1, x_i \mid z_0) \tag{6}$$

where $\mathrm{ACDE}(x) \equiv E[Y_{x1} - Y_{x0}]$. We may also obtain bounds on average controlled direct effects for AT and NT:

$$1 - \frac{p(y_0, x_0|z_0) + p(y_1, x_0|z_1)}{p(x_0|z_0) - p(x_1|z_1)} \leq \mathrm{ACDE}_{\mathrm{NT}}(x_0) \leq \frac{p(y_0, x_0|z_1) + p(y_1, x_0|z_0)}{p(x_0|z_0) - p(x_1|z_1)} - 1,$$

$$1 - \frac{p(y_0, x_1|z_1) + p(y_1, x_1|z_0)}{p(x_1|z_1) - p(x_0|z_0)} \leq \mathrm{ACDE}_{\mathrm{AT}}(x_1) \leq \frac{p(y_0, x_1|z_0) + p(y_1, x_1|z_1)}{p(x_1|z_1) - p(x_0|z_0)} - 1,$$

where $\mathrm{ACDE}_{\mathfrak{t}_X}(x) \equiv E[Y_{x1} - Y_{x0} \mid \mathfrak{t}_X]$; ($\mathrm{ACDE}_{\mathrm{NT}}(x_0)$ and $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$ are also referred to as 'principal stratum direct effects' for $X = 0$ and $X = 1$ respectively). $\mathrm{ACDE}(x_0)$ may be bounded away from 0 iff $\mathrm{ACDE}_{\mathrm{NT}}(x_0)$ may be bounded away from 0 in the same direction (hence Ex$_{\mathrm{NT}}$ does not hold). Likewise with $\mathrm{ACDE}(x_1)$, $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$ and Ex$_{\mathrm{AT}}$. Note that since any distribution $p(y, x|z)$ may violate at most one of the four inequalities (4) and (5), in the absence of further assumptions (such as Mon$_X$), every distribution is either compatible with Ex$_{\mathrm{AT}}$ or Ex$_{\mathrm{NT}}$ (or both).

It may be shown that the model imposing Ex$_{\mathrm{NT}}$ alone is characterized by (4), while the model imposing Ex$_{\mathrm{AT}}$ is given by (5); see Richardson and Robins (2010).

Inequalities defining models without Defiers

The assumption Mon_X, that there are no Defiers, implies:

$$p(x_1 \mid z_1) \geq p(x_1 \mid z_0) \tag{7}$$

since the left and right sides are the proportions of (AT or CO) and AT respectively. Thus (7) characterizes the observed distributions resulting from Mon_X alone.

Results of Balke and Pearl (1997) imply that the model assuming Mon_X + $\mathrm{Ex}_{\mathrm{NT}}$ + $\mathrm{Ex}_{\mathrm{AT}}$ implies the following inequalities:

$$p(y_1, x_0 \mid z_1) \;\leq\; p(y_1, x_0 \mid z_0), \quad p(y_0, x_0 \mid z_1) \;\leq\; p(y_0, x_0 \mid z_0), \tag{8}$$

$$p(y_1, x_1 \mid z_1) \;\geq\; p(y_1, x_1 \mid z_0), \quad p(y_0, x_1 \mid z_1) \;\geq\; p(y_0, x_1 \mid z_0). \tag{9}$$

A distribution $p(y, x \mid z)$ may violate all of these simultaneously. These inequalities are invariant to relabelling Y, and to relabelling X and Z simultaneously, but not individually; this is not surprising since relabelling X or Z alone will turn Defiers into Compliers and vice-versa. The inequalities (8) and (9) imply (7), (4) and (5).

It may be shown that (8) and (9) characterize the set of distributions $p(y, x|z)$ arising from the potential outcomes model Mon_X + $\mathrm{Ex}_{\mathrm{NT}}$ + $\mathrm{Ex}_{\mathrm{AT}}$. Likewise, the model imposing Mon_X + $\mathrm{Ex}_{\mathrm{NT}}$ is characterized by (7) and (8), while Mon_X + $\mathrm{Ex}_{\mathrm{AT}}$ is given by (7) and (9).

An interpretation of (8) and (9) is given by the following lower bound on π_{DE} in the model that imposes $\mathrm{Ex}_{\mathrm{NT}}$ + $\mathrm{Ex}_{\mathrm{AT}}$ (but not Mon_X):

$$\pi_{\mathrm{DE}} \geq \max \left\{ \begin{array}{ll} 0, & p(x_1 \mid z_0) - p(x_1 \mid z_1), \\ p(y_1, x_0|z_1) - p(y_1, x_0|z_0), & p(y_0, x_0|z_1) - p(y_0, x_0|z_0), \\ p(y_1, x_1|z_0) - p(y_1, x_1|z_1), & p(y_0, x_1|z_0) - p(y_0, x_1|z_1) \end{array} \right\}; \tag{10}$$

see Richardson and Robins (2010). Requiring that the lower bound be zero, as required by Mon_X, leads directly to the inequalities (7), (8) and (9).

Another interpretation of (8) and (9) arises in the model Mon_X that (solely) assumes that there are no Defiers. Under Mon_X we may obtain tighter bounds on the ACDE for AT and NT:

$$\begin{aligned} p(y_1|x_0, z_1) - \min\left(p(y_1, x_0|z_0)/p(x_0|z_1), 1\right) \leq \mathrm{ACDE}_{\mathrm{NT}}(x_0) \leq \\ p(y_1|x_0, z_1) - \max\left(0, 1 - (p(y_0, x_0|z_0)/p(x_0|z_1))\right), \end{aligned} \tag{11}$$

$$\begin{aligned} \max\left(0, 1 - (p(y_0, x_1|z_1)/p(x_1|z_0))\right) - p(y_1|x_1, z_0) \leq \mathrm{ACDE}_{\mathrm{AT}}(x_1) \leq \\ \min\left(p(y_1, x_1|z_1)/p(x_1|z_0), 1\right) - p(y_1|x_1, z_0). \end{aligned} \tag{12}$$

However, the bounds (6) on the global $\mathrm{ACDE}(x_i)$ remain sharp, being unchanged by the assumption of monotonicity.

It is simple to show that $\mathrm{ACDE}_{\mathrm{NT}}(x_0)$ is bounded away from 0 by (11) iff one of the inequalities (8) is violated; likewise for $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$, (12) and (9). Thus if Mon_X, and hence (7) holds, then at most one inequality in each of the pairs (8) and (9) may be violated. However, in contrast to the case without the monotonicity assumption, since it is possible for a distribution $p(y, x|z)$ to violate one inequality in each pair simultaneously, ACDE_{NT} and ACDE_{AT} may *both* be bounded away from zero. Thus under the assumption of No Defiers both Ex_{NT} and Ex_{AT} may be inconsistent with $p(y, x|z)$.

Table 4 summarizes the constraints for the eight models we consider. For frequentist approaches to testing these constraints see Ramsahai (2008).

Table 4: *Models and implied sets of distributions for* $p(y, x|z)$*;* (8) *and* (9) *imply* (7).

Model	Assumptions	Constraints on $p(y, x\|z)$
Saturated	Randomization (1)	None
Ex_{NT}	(1), Exclusion for NT	(4)
Ex_{AT}	(1), Exclusion for AT	(5)
$\text{Ex}_{\text{AT}} + \text{Ex}_{\text{NT}}$	(1), Exclusion for AT and NT	(4), (5)
Mon_X	(1), No Defiers	(7)
$\text{Mon}_X + \text{Ex}_{\text{NT}}$	(1), No Defiers, Exclusion for NT	(7), (8)
$\text{Mon}_X + \text{Ex}_{\text{AT}}$	(1), No Defiers, Exclusion for AT	(7), (9)
Mon_X $+ \text{Ex}_{\text{NT}} + \text{Ex}_{\text{AT}}$	(1), No Defiers, Exclusion for NT and AT	[(7)], (8), (9),

Table 5: *Summary of Flu Vaccine Data; originally from McDonald et al. (1992); analyzed by Hirano et al. (2000).*

z	x	y	count	$p(y, x\|z_0)$	z	x	y	count	$p(y, x\|z_1)$
0	0	0	99	0.071	1	0	0	84	0.057
0	0	1	1027	0.739	1	0	1	935	0.635
0	1	0	30	0.022	1	1	0	31	0.021
0	1	1	233	0.168	1	1	1	422	0.287
			1389					1472	

ANALYSIS OF FLU VACCINE DATA

We consider the influenza vaccine data from McDonald et al. (1992) which was previously analyzed by Hirano et al. (2000); see Table 5. Here the instrument Z was whether a patient's physician was sent a card asking them to remind patients to obtain flu shots, or not; X is whether or not the patient did in fact get a flu shot. Finally $Y = 1$ indicates that a patient was *not* hospitalized.

To examine the support for the restrictions on $p(y, x|z)$ we fitted a saturated model with uniform priors and then evaluated the posterior probability that the inequalities (4), (5), (7), (8) and (9) are violated. For a model without covariates these probabilities are shown in the first line of Table 6. The posterior probability that at least one of the inequalities (9) fails to hold has probability greater than 0.5; a similar conclusion may be arrived at by inspection of the row of Table 5 for $(y = 0, x = 1)$. If (9) is violated then, under the assumptions of no Defiers (which seems plausible) and randomization, there is a direct effect for Always Takers.

Hirano et al. (2000) place priors over the (partially) identified parameters of the potential outcome model and compute posteriors for the Intent-To-Treat effect:

$$\text{ITT}_{\mathfrak{t}_X} \equiv E[Y_{X_{z_1}1} - Y_{X_{z_0}0} \mid \mathfrak{t}_X]$$

for NT, AT and CO under the models Mon_X, $\text{Mon}_X+\text{Ex}_{\text{AT}}$, $\text{Mon}_X+\text{Ex}_{\text{NT}}$ and $\text{Mon}_X+\text{Ex}_{\text{AT}}+\text{Ex}_{\text{NT}}$. Under an additional extra exclusion assumption for compliers, $\gamma^{00}_{\text{CO}} = \gamma^{01}_{\text{CO}}$, ITT_{CO} is equal to the Complier Average Causal Effect of X on Y, $\text{CACE}_{\text{CO}} \equiv E[Y_{X_1} - Y_{X_0} \mid \mathfrak{t}_X] = \gamma^{1\cdot}_{\text{CO}} - \gamma^{0\cdot}_{\text{CO}}$.

In Figure 6 we display the joint posterior distributions over upper and lower bounds on ITT_{CO} under each of the eight models we consider. (Each scatterplot is based on 2000 simulations.) The bounds were computed by applying the methods described in §§2–3 of Richardson and Robins (2010).

Table 6: *Posterior probabilities that inequalities are violated under models that do not impose constraints. The two models without Age used a uniform prior on* $\Delta_3 \times \Delta_3$*; that with Age used logistic regressions with Normal priors. Columns* (4), (5), (8) *and* (9) *give the probability that at least one inequality is violated;* (8)+(9) *is the probability of at least one violation in each pair.*

age	copd	(4)	(5)	(7)	(8)	(9)	(8)+(9)	(8) both	(9) both
-	-	0	0	0	0.0603	0.5411	0.0343	0	0
-	N	0	0	0	0.0704	0.4635	0.0347	0	0
-	Y	0	0	0.0014	0.2969	0.5865	0.1829	0.0003	0.0003
60	N	0	0	0	0.0768	0.2600	0.0306	0	0
60	Y	0	0	0.0064	0.3016	0.6222	0.2074	0.0014	0.0016
70	N	0	0	0	0.0422	0.5958	0.0288	0	0
70	Y	0	0	0.0080	0.4154	0.5580	0.2626	0.0026	0.0030
80	N	0	0	0.0002	0.0900	0.8064	0.0764	0	0
80	Y	0	0	0.0608	0.5338	0.5320	0.3214	0.0116	0.0128

5. INCORPORATING COVARIATES

In many situations we wish to examine causal effects in sub-populations defined by baseline covariates V. In this situation we assume that the randomization assumption (1), and (when we impose them) Mon_X, Ex_{AT}, and Ex_{NT} hold within levels of V. With discrete covariates taking a small number of levels we may simply repeat our analysis within each level of V. However in order to incorporate continuous baseline covariates we require a parametrization of each of the sets of distributions appearing in Table 4. For each model we provide a smooth variation independent parametrization of the relevant subset of $\Delta_3 \times \Delta_3$. This allows us to construct (multivariate) generalized linear models for $p(y, x|z)$ as a function of V.

Parametrization of Models with Defiers

Consider first the set of distributions $p(y, x|z)$ that result from models assuming both Ex_{AT} and Ex_{NT}, and hence satisfy the inequalities (4) and (5). It is clear that for any distribution $p(y, x|z_0)$ there exists a distribution $p(y, x|z_1)$ such that the pair satisfy (4) and (5). Thus the set of distributions obeying (4) and (5) is:

$$\left\{ p(y,x|z) \;\middle|\; p(y,x|z_0) \in \Delta_3,\;\; p(y,x|z_1) \in \Delta_3 \cap \bigcap_{i,j\in\{0,1\}} H_{ij}(p(y_{1-i},x_j|z_0)) \right\} \qquad (13)$$

where $H_{ij}(p(y_{1-i}, x_j|z_0)) \equiv \{p(y, x|z_1) \mid p(y_i, x_j|z_1) \leq 1 - p(y_{1-i}, x_j|z_0)\}$, i.e. a half-space. We parametrize the set (13) via the unrestricted distributions $p(y, x|z_0)$,

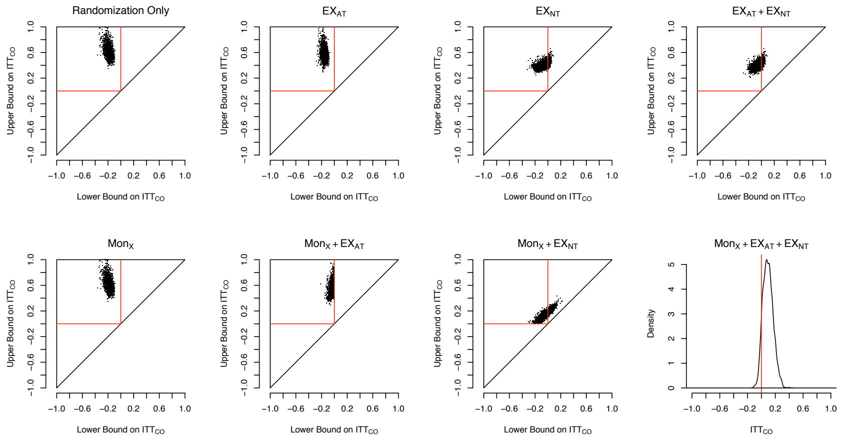


Figure 6: *Posterior distributions for upper and lower bounds on* $\mathrm{ITT}_{\mathrm{CO}}$*; under* Mon_X+$\mathrm{Ex}_{\mathrm{AT}}$+$\mathrm{Ex}_{\mathrm{NT}}$ *the parameter is identified.*

$p(x|z_1)$ and two further parameters, ψ_0, ψ_1 where

$$\psi_i \;\equiv\; \log\left(\frac{p(y_0, x_i|z_1)(1 - p(y_1, x_i|z_1) - \{p(y_0, x_i|z_0)\})}{p(y_1, x_i|z_1)(1 - p(y_0, x_i|z_1) - \{p(y_1, x_i|z_0)\})}\right). \tag{14}$$

The inverse map from $(p(y, x|z_0), p(x|z_1), \psi_0, \psi_1)$ to $p(y, x|z_1)$ is given by:

$$\begin{aligned} p(y_1, x_i|z_1) \;&=\; \left(-b_i + \sqrt{b_i^2 + 4(e^{\psi_i} - 1)p(x_i|z_1)(1 - p(y_0, x_i|z_0))}\right) \Big/ \left(2(e^{\psi_i} - 1)\right), \\ p(y_0, x_i|z_1) \;&=\; p(x_i|z_1) - p(y_1, x_i|z_1), \end{aligned}$$

for $i = 0, 1$, where $b_i \;=\; e^{\psi_i}(p(x_{1-i}|z_1) - p(y_1, x_i|z_0)) + p(x_i|z_1) + 1 - p(y_0, x_i|z_0)$.

If we let

$$\tilde{\psi}_i \;\equiv\; \log\left(\frac{p(y_0, x_i|z_1)(1 - p(y_1, x_i|z_1))}{p(y_1, x_i|z_1)(1 - p(y_0, x_i|z_1))}\right), \tag{15}$$

the parameter defined by removing the terms in braces from (14), then the model imposing $\mathrm{Ex}_{\mathrm{AT}}$ alone may be parametrized via $(p(y, x|z_0), p(x|z_1), \tilde{\psi}_0, \psi_1)$. Similarly $(p(y, x|z_0), p(x|z_1), \psi_0, \tilde{\psi}_1)$ parametrizes the model imposing $\mathrm{Ex}_{\mathrm{NT}}$ alone.

Inverse maps for these models are similar to that for $\mathrm{Ex}_{\mathrm{AT}}$ + $\mathrm{Ex}_{\mathrm{NT}}$.

Parameterization of Models without Defiers

The model with Mon_X alone may be parametrized via $p(y, x|z_0)$, $\nu_{x|z_1}$ and $p(y|x_1, z_0)$, where

$$\nu_{x|z_1} \;\equiv\; \mathrm{logit}(p(x_0|z_1)/p(x_0|z_0)).$$

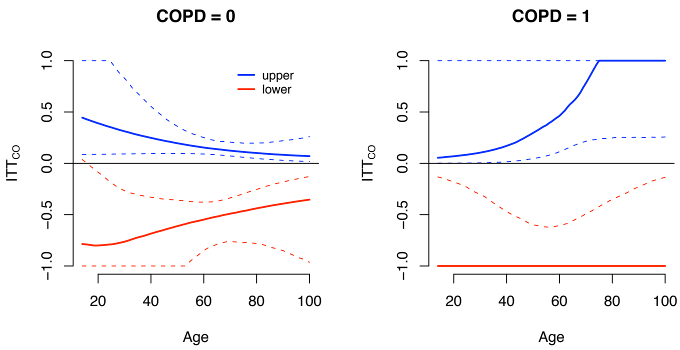


Figure 7: *Posteriors for* $\mathrm{ITT}_{\mathrm{CO}}$ *under the* Mon_X *model which precludes Defiers, as a function of Age; (left) without COPD; (right) with COPD. Dashed lines indicate* 2.5 *and* 97.5 *percentiles for the upper and lower bounds.* $p(y, x|z_0, v)$ *was parametrized via logits* $\theta_{y|x_0,z_0}$, $\theta_{y|x_1,z_0}$ *and* $\theta_{x|z_0}$. *These logits and* $\nu_{x|z_1}$, ϕ_0 *and* φ_1 *are modelled as linear functions of Age and COPD.*

The model $\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{NT}} + \mathrm{Ex}_{\mathrm{AT}}$, may be parametrized via $p(y, x|z_0)$, $\nu_{x|z_1}$, ϕ_0 and φ_1 where the latter are defined via:

$$\begin{aligned} \phi_0 &\equiv \log\left(\frac{p(y_0, x_0|z_1)(1 - p(y_1, x_0|z_1) - \{1 - p(y_1, x_0|z_0)\})}{p(y_1, x_0|z_1)(1 - p(y_0, x_0|z_1) - \{1 - p(y_0, x_0|z_0)\})}\right), \\ \varphi_1 &\equiv \log\left(\frac{(1 - p(y_1, x_1|z_1))(p(y_0, x_1|z_1) - \{p(y_0, x_1|z_0)\})}{(1 - p(y_0, x_1|z_1))(p(y_1, x_1|z_1) - \{p(y_1, x_1|z_0)\})}\right). \end{aligned}$$

The inverse map from $(p(y, x|z_0), \nu_{x|z_1}, \phi_0, \varphi_1)$ to $p(y, x|z)$ is given by:

$$\begin{aligned} p(x_0|z_1) &= p(x_0|z_0)\,\mathrm{expit}\,\nu_{x|z_1}, \\ p(y_1, x_0|z_1) &= \left(-c_0 + \sqrt{c_0^2 + 4(e^{\phi_0} - 1)p(x_0|z_1)p(y_1, x_0|z_0)}\right)\Big/\left(2(e^{\phi_0} - 1)\right), \\ p(y_0, x_0|z_1) &= p(x_0|z_1) - p(y_1, x_0|z_1), \\ p(y_0, x_1|z_1) &= 1 - \left(\frac{-c_1 + \sqrt{c_1^2 + 4(e^{\varphi_1} - 1)(1 + p(x_0|z_1))(1 - p(y_0, x_1|z_0))}}{2(e^{\varphi_1} - 1)}\right), \\ p(y_1, x_1|z_1) &= 1 - p(x_0|z_1) - p(y_0, x_1|z_1), \end{aligned}$$

where

$$\begin{aligned} c_0 &= e^{\phi_0}(p(y_0, x_0|z_0) - p(x_0|z_1)) + p(y_1, x_0|z_0) + p(x_0|z_1), \\ c_1 &= 1 - e^{\varphi_1}(p(y_1, x_1|z_0) + p(x_0|z_1)) + 1 - p(y_0, x_1|z_0) + p(x_0|z_1). \end{aligned}$$

$\text{Mon}_X + \text{Ex}_{\text{AT}}$ may be parametrized via $p(y, x|z_0)$, $\nu_{x|z_1}$, $\tilde{\phi}_0$ and φ_1 where

$$\tilde{\phi}_0 \equiv \log\left(\frac{p(y_0, x_0|z_1)(1 - p(y_1, x_0|z_1))}{p(y_1, x_0|z_1)(1 - p(y_0, x_0|z_1))}\right)$$

simply omits the terms in braces in ϕ_0.

$\text{Mon}_X + \text{Ex}_{\text{NT}}$ may be parametrized via $p(y, x|z_0)$, $\nu_{x|z_1}$, ϕ_0 and $\tilde{\varphi}_1$ where

$$\tilde{\varphi}_1 \equiv \log\left(\frac{(1 - p(y_1, x_1|z_1))p(y_0, x_1|z_1)}{(1 - p(y_0, x_1|z_1))p(y_1, x_1|z_1)}\right)$$

again simply omits the terms in braces in φ_1.

Inverse maps for these models are similar to that for $\text{Mon}_X + \text{Ex}_{\text{NT}} + \text{Ex}_{\text{AT}}$.

Table 7: *Parameterization of Models. Distributions appearing in the parameter list are unrestricted.*

Model	Parameters
Saturated	$p(x, y\|z)$
Ex_{NT}	$p(x, y\|z_0)$, $p(x\|z_1)$, ψ_0, $\tilde{\psi}_1$
Ex_{AT}	$p(x, y\|z_0)$, $p(x\|z_1)$, $\tilde{\psi}_0$, ψ_1
$\text{Ex}_{\text{AT}} + \text{Ex}_{\text{NT}}$	$p(x, y\|z_0)$, $p(x\|z_1)$, ψ_0, ψ_1
Mon_X	$p(x, y\|z_0)$, $\nu_{x\|z_1}$, $p(y\|x, z_1)$
$\text{Mon}_X + \text{Ex}_{\text{NT}}$	$p(x, y\|z_0)$, $\nu_{x\|z_1}$, ϕ_0, $\tilde{\varphi}_1$
$\text{Mon}_X + \text{Ex}_{\text{AT}}$	$p(x, y\|z_0)$, $\nu_{x\|z_1}$, $\tilde{\phi}_0$, φ_1
Mon_X $+ \text{Ex}_{\text{NT}} + \text{Ex}_{\text{AT}}$	$p(x, y\|z_0)$, $\nu_{x\|z_1}$, ϕ_0, φ_1

Flu Vaccine Data Revisited

Following the analysis of Hirano et al. (2000) we consider the baseline covariates Age, and COPD (chronic obstructive pulmonary disease). Table 6 shows the posterior probability of violations of constraints under saturated models stratifying on COPD, and under a model specified via 6 logistic regressions (for $p(x|z)$ and $p(y|x, z)$) each with intercept, Age, COPD and COPD×Age.

Finally, to illustrate our parametrization we fitted the model Mon_X. Figure 7 shows posterior distributions on bounds for ITT_{CO} under Mon_X. The model was specified via logistic regressions for $p(y|x_0, z)$, $p(x|z_0)$ and generalized linear models for $\nu_{x|z_1}$, ϕ_0 and φ_1, again each with intercept, Age, COPD and COPD×Age. Diffuse independent Normal priors were used. Sampling was performed via a Metropolis algorithm. The proposal for each of the six GLMs was multivariate normal, mean 0, covariance matrix $\hat{\sigma}_k^2 \mathbf{V}^T \mathbf{V}$ where $\mathbf{V}$ is the $n \times 4$ model matrix, and $\hat{\sigma}_k^2$ $(k = 1, \ldots, 6)$ is an estimate of the variance of the specific parameter, obtained via the delta method at the empirical MLE for $p(y, x|z)$. There were 2000 burn-in iterations followed by 5000 main iterations. The Markov chain was initialized by setting all of the generalized linear model parameters to 0.

Acknowledgements

This research was supported by the U.S. National Science Foundation (CRI 0855230) and U.S. National Institutes of Health (R01 AI032475). We thank McDonald, Hiu and Tierney for giving us permission to use their flu vaccine data. Thanks also to Susan Shortreed who implemented part of the Gibbs sampler used in §3, and to Adrian Raftery for helpful suggestions.

REFERENCES

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

Bonet, B. (2001). Instrumentality tests revisited. In *Proceedings of the 17*$^{\text{th}}$ *Conference on Uncertainty in Artificial Intelligence*, pp. 48–55.

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics 64*, 695–701.

Chickering, D. and J. Pearl (1996). A clinician's tool for analyzing non-compliance. In *AAAI-96 Proceedings*, pp. 1269–1276.

Efron, B. and D. Feldman (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc. 86*, 9–26.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc 168*, 267–306.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science 20*(2), 111–140.

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

Leamer, E. (1978). *Specification Searches*. New York: Wiley.

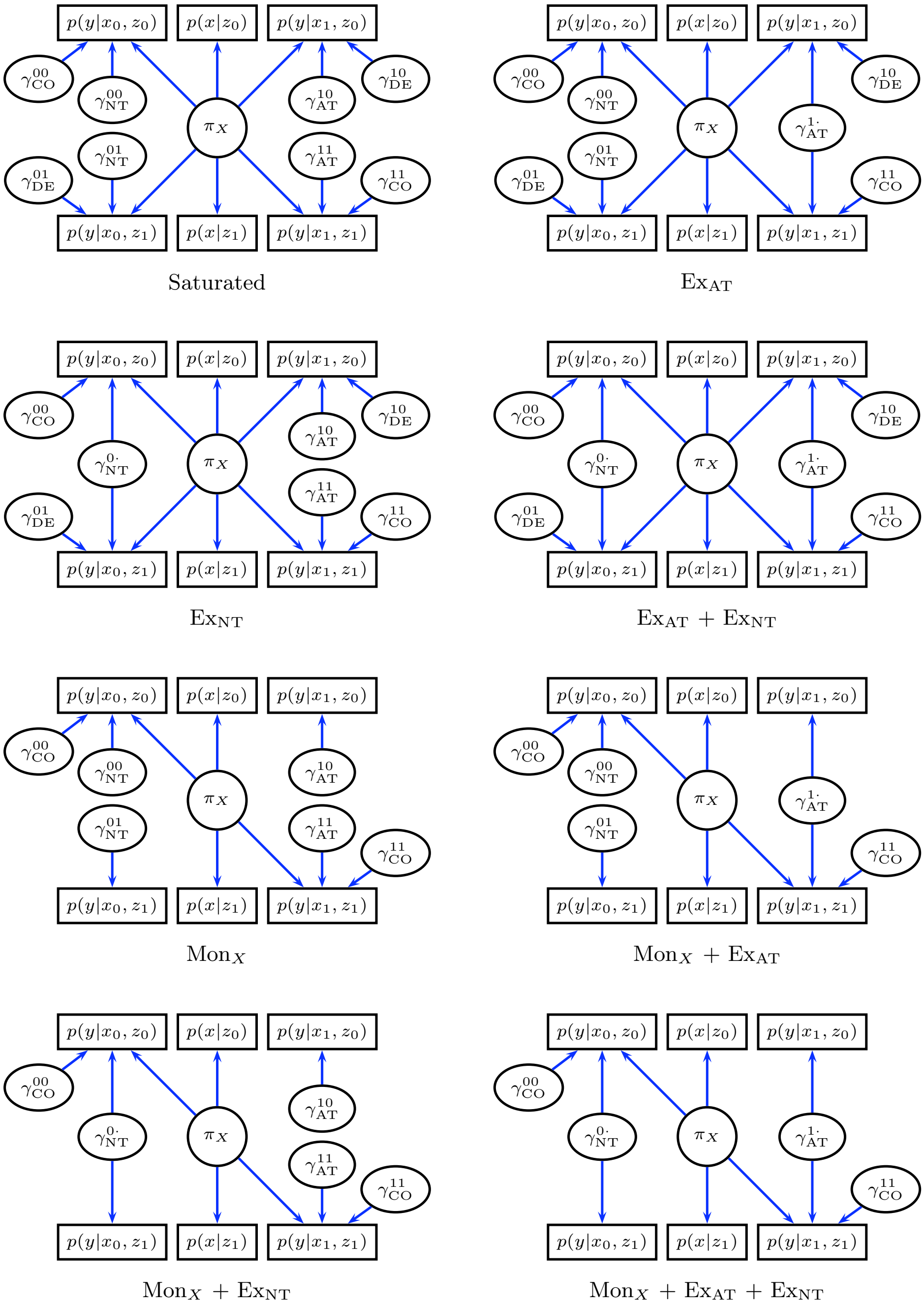
McDonald, C., S. Hiu, and W. Tierney (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing 9*, 304–312.

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

Ramsahai, R. (2008). *Causal Inference with Instruments and Other Supplementary Variables*. Ph. D. thesis, University of Oxford, Oxford, UK.

Richardson, T. S. and J. M. Robins (2010). Analysis of the binary instrumental variable model. In R. Dechter, H. Geffner, and J. Halpern (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, Chapter 25, pp. 415–444. London: College Publications.

Robins, J. M., T. S. Richardson, and P. Spirtes (2009). Identification and inference for direct effects. Technical Report 563, Dept. of Statistics, Univ. Washington.


$p(y|x_0, z_0)$
$p(x|z_0)$
$p(y|x_1, z_0)$
$\gamma^{00}_{\mathrm{CO}}$
$\gamma^{00}_{\mathrm{NT}}$
$\gamma^{01}_{\mathrm{NT}}$
$\gamma^{0\cdot}_{\mathrm{NT}}$
$\gamma^{01}_{\mathrm{DE}}$
π_X
$\gamma^{10}_{\mathrm{AT}}$
$\gamma^{11}_{\mathrm{AT}}$
$\gamma^{1\cdot}_{\mathrm{AT}}$
$\gamma^{10}_{\mathrm{DE}}$
$\gamma^{11}_{\mathrm{CO}}$
$p(y|x_0, z_1)$
$p(x|z_1)$
$p(y|x_1, z_1)$

Saturated

$\mathrm{Ex}_{\mathrm{AT}}$

$\mathrm{Ex}_{\mathrm{NT}}$

$\mathrm{Ex}_{\mathrm{AT}} + \mathrm{Ex}_{\mathrm{NT}}$

Mon_X

$\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{AT}}$

$\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{NT}}$

$\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{AT}} + \mathrm{Ex}_{\mathrm{NT}}$

Figure 8: Functional dependencies in the eight models. Terms $\gamma^{ij}_{\mathsf{t}_X}$ that do not appear in the likelihood are not shown. See also Table 4.