Analysis of the Binary Instrumental Variable Model

Thomas S. Richardson
University of Washington

James M. Robins
Harvard School of Public Health

Working Paper no. 99
Center for Statistics and the Social Sciences
University of Washington

29 March, 2010

Abstract

We give an explicit geometric characterization of the set of distributions over counterfactuals that are compatible with a given observed joint distribution for the observables in the binary instrumental variable model.

This paper will appear as Chapter 25 in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. R. Dechter, H. Geffner and J.Y. Halpern, Editors, College Publications, UK.

1 Introduction

Pearl's seminal work on instrumental variables [Chickering and Pearl 1996; Balke and Pearl 1997] for discrete data represented a leap forwards in terms of understanding: Pearl showed that, contrary to what many had supposed based on linear models, in the discrete case the assumption that a variable was an instrument could be subjected to empirical test. In addition, Pearl improved on earlier bounds [Robins 1989] for the average causal effect (ACE) in the absence of any monotonicity assumptions. Pearl's approach was also innovative insofar as he employed a computer algebra system to derive analytic expressions for the upper and lower bounds.

In this paper we build on and extend Pearl's work in two ways. First we show the geometry underlying Pearl's bounds. As a consequence we are able to derive bounds on the average causal effect for all four compliance types. Our analysis also makes it possible to perform a sensitivity analysis using the distribution over compliance types. Second our analysis provides a clear geometric picture of the instrumental inequalities, and allows us to isolate the counterfactual assumptions necessary for deriving these tests. This may be seen as analogous to the geometric study of models for two-way tables [Fienberg and Gilbert 1970; Erosheva 2005]. Among other things this allows us to clarify which are the alternative hypotheses against which Pearl's test has power. We also relate these tests to recent work of Pearl's on bounding direct effects [Cai, Kuroki, Pearl, and Tian 2008].

2 Background

We consider three binary variables, X, Y and Z. Where:

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

X is the treatment received;

Y is the response.

For X and Z, we will use 0 to indicate placebo, and 1 to indicate drug. For Y we take 1 to indicate a desirable outcome, such as survival. X_z is the treatment a patient would receive if assigned to $Z = z$. We follow convention by referring to the four *compliance* types:

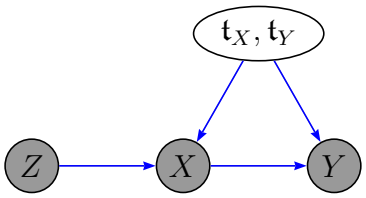

Figure 1: Graphical representation of the IV model given by assumptions (1) and (2).The shaded nodes are observed.

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

Since we suppose the counterfactuals are well-defined, if $Z = z$ then $X = X_z$. Similarly we consider counterfactuals Y_{xz} for Y. Except where explicitly noted we will make the exclusion restrictions:

$$Y_{x=0,z=0} = Y_{x=0,z=1} \qquad Y_{x=1,z=0} = Y_{x=1,z=1} \tag{1}$$

for each patient, so that a patient's outcome only depends on treatment assigned via the treatment received. One consequence of the analysis below is that these equations may be tested separately. We may thus similarly enumerate four types of patient in terms of their *response* to received treatment:

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

As before, it is implicit in our notation that if $X = x$, then $Y_x = Y$; this is referred to as the 'consistency assumption' (or axiom) by Pearl among others. In what follows we will use $\mathfrak{t}_X$ to denote a generic compliance type in the set $\mathbb{D}_X$, and $\mathfrak{t}_Y$ to denote a generic response type in the set $\mathbb{D}_Y$. We thus have 16 patient types:

$$\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle \in \{\text{NT, CO, DE, AT}\} \times \{\textit{NR, HE, HU, AR}\} \equiv \mathbb{D}_X \times \mathbb{D}_Y \equiv \mathbb{D}.$$

(Here and elsewhere we use angle brackets $\langle \mathfrak{t}_X, \mathfrak{t}_Y \rangle$ to indicate an ordered pair.) 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 let

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

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

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

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

Except where explicitly noted we will make the randomization assumption that the distribution of types $\langle \mathbf{t}_X, \mathbf{t}_Y \rangle$ is the same in both arms:

$$Z \perp\!\!\!\perp \{X_{z=0}, X_{z=1}, Y_{x=0}, Y_{x=1}\}. \tag{2}$$

A graph corresponding to the model given by (1) and (2) is shown in Figure 1.

2.0.1 Notation

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

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

There are several simple geometric constructions that we will use repeatedly. In consequence we introduce these in a generic setting.

2.1 Joints compatible with fixed margins

Consider a bivariate random variable $U = \langle U_1, U_2 \rangle \in \{0, 1\} \times \{0, 1\}$. Now for fixed $c_1, c_2 \in [0, 1]$ consider the set

$$\mathcal{P}_{c_1,c_2} = \left\{ p \;\middle|\; \sum_{u_2} p(1, u_2) = c_1 \;;\; \sum_{u_1} p(u_1, 1) = c_2 \right\}$$

in other words, $\mathcal{P}_{c_1,c_2}$ is the set of joint distributions on U compatible with fixed margins $p(U_i = 1) = c_i$, $i = 1, 2$.

It is not hard to see that $\mathcal{P}_{c_1,c_2}$ is a one-dimensional subset (line segment) of the 3-dimensional simplex of distributions for U. We may describe it explicitly as follows:

$$\left\{ \begin{array}{lcl} p(1,1) &=& t \\ p(1,0) &=& c_1 - t \\ p(0,1) &=& c_2 - t \\ p(0,0) &=& 1 - c_1 - c_2 + t \end{array} \qquad t \in \left[\max\left\{0, (c_1 + c_2) - 1\right\}, \min\left\{c_1, c_2\right\}\right] \right\}. \tag{3}$$

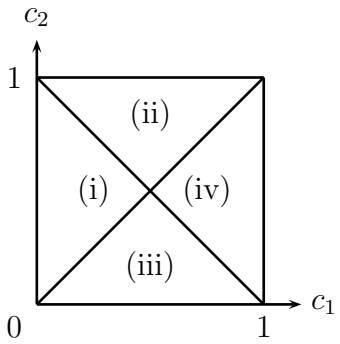

Figure 2: The four regions corresponding to different supports for t in (3); see Table 1.

See also [Pearl 2000] Theorem 9.2.10. The range of t, or equivalently the support for $p(1,1)$, is one of four intervals, as shown in Table 1. These cases correspond to the four regions show

	$c_1 \leq 1 - c_2$		$c_1 \geq 1 - c_2$	
$c_1 \leq c_2$	(i)	$t \in [0, c_1]$	(ii)	$t \in [c_1 + c_2 - 1, c_1]$
$c_1 \geq c_2$	(iii)	$t \in [0, c_2]$	(iv)	$t \in [c_1 + c_2 - 1, c_2]$

Table 1: The support for t in (3) in each of the four cases relating c_1 and c_2.

in Figure 2.

Finally, we note that since for $c_1, c_2 \in [0,1]$, $\max\{0, (c_1 + c_2) - 1\} \leq \min\{c_1, c_2\}$, it follows that $\{\langle c_1, c_2\rangle \mid \mathcal{P}_{c_1,c_2} \neq \emptyset\} = [0,1]^2$. Thus for every pair of values $\langle c_1, c_2\rangle$ there exists a joint distribution $p(U_1, U_2)$ for which $p(U_i = 1) = c_i$, $i = 1, 2$.

2.2 Two quantities with a specified average

We now consider the set:

$$\mathcal{Q}_{c,\alpha} = \{\langle u, v\rangle \mid \alpha u + (1-\alpha)v = c,\ u, v \in [0,1]\}$$

where $c, \alpha \in [0,1]$. In words, $\mathcal{Q}_{c,\alpha}$ is the set of pairs of values $\langle u, v\rangle$ in $[0,1]$ which are such that the weighted average $\alpha u + (1-\alpha)v$ is c.

It is simple to see that this describes a line segment in the unit square. Further consideration shows that for any value of $\alpha \in [0,1]$, the segment will pass through the point $\langle c, c\rangle$ and will be contained within the union of two rectangles:

$$([c,1] \times [0,c]) \cup ([0,c] \times [1,c]).$$

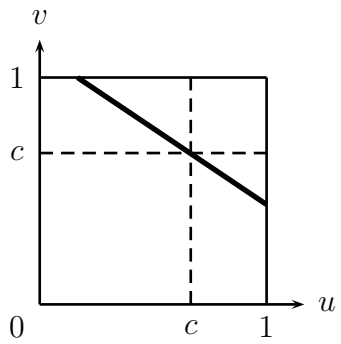

Figure 3: Illustration of $\mathcal{Q}_{c,\alpha}$.

The slope of the line is negative for $\alpha \in (0,1)$. For $\alpha \in (0,1)$ the line segment may be parametrized as follows:

$$\left\{ \begin{array}{rcll} u & = & (c - t(1-\alpha))/\alpha, & t \in \left[\max\left(0, \dfrac{c-\alpha}{1-\alpha}\right), \min\left(\dfrac{c}{1-\alpha}, 1\right) \right] \\ v & = & t, & \end{array} \right\}.$$

The left and right endpoints of the line segment are:

$$\langle u, v \rangle = \left\langle \max\big(0, 1 + (c-1)/\alpha\big)\,,\ \min\big(c/(1-\alpha), 1\big) \right\rangle$$

and

$$\langle u, v \rangle = \left\langle \min\big(c/\alpha, 1\big)\,,\ \max\big(0, (c-\alpha)/(1-\alpha)\big) \right\rangle$$

respectively. See Figure 3.

2.3 Three quantities with two averages specified

We now extend the discussion in the previous section to consider the set:

$$\begin{aligned} \mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} = \{ \langle u, v, w \rangle \mid\ & \alpha_1 u + (1-\alpha_1) w = c_1, \\ & \alpha_2 v + (1-\alpha_2) w = c_2,\ u, v, w \in [0,1] \}\,. \end{aligned}$$

In words, this consists of the set of triples $\langle u, v, w \rangle \in [0,1]^3$ for which pre-specified averages of u and w (via α_1), and v and w (via α_2) are equal to c_1 and c_2 respectively.

If this set is not empty, it is a line segment in $[0,1]^3$ obtained by the intersection of two rectangles:

$$\Big(\{ \langle u, w \rangle \in \mathcal{Q}_{c_1,\alpha_1} \} \times \{ v \in [0,1] \} \Big) \cap \Big(\{ \langle v, w \rangle \in \mathcal{Q}_{c_2,\alpha_2} \} \times \{ u \in [0,1] \} \Big); \qquad (4)$$

see Figures 4 and 5. For $\alpha_1, \alpha_2 \in (0,1)$ we may parametrize the line segment (4) as follows:

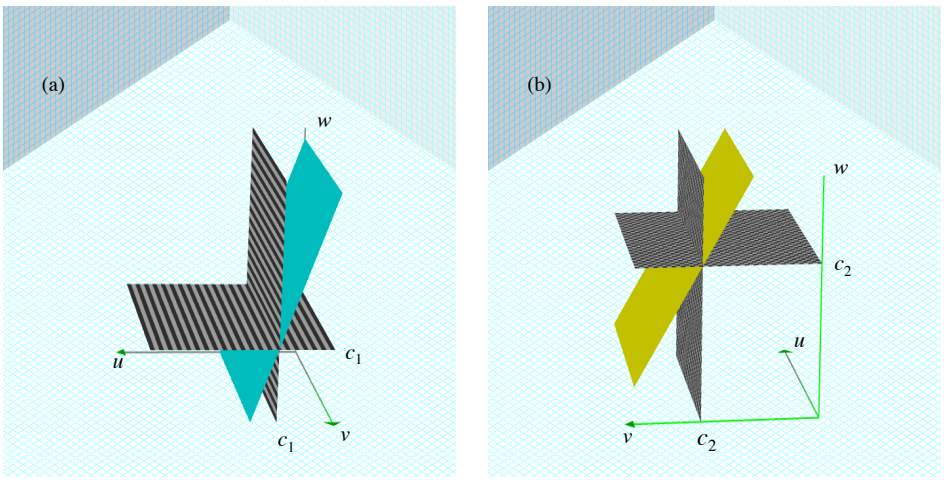

Figure 4: (a) The plane without stripes is $\alpha_1 u + (1-\alpha_1)w = c_1$. (b) The plane without checks is $\alpha_2 v + (1-\alpha_2)w = c_2$.

$$\left\{ \begin{array}{lcll} u & = & (c_1 - t(1-\alpha_1))/\alpha_1, & \\ v & = & (c_2 - t(1-\alpha_2))/\alpha_2, & t \in [t_l, t_u] \\ w & = & t, & \end{array} \right\},$$

where

$$t_l \equiv \max\left\{0, \frac{c_1-\alpha_1}{1-\alpha_1}, \frac{c_2-\alpha_2}{1-\alpha_2}\right\}, \qquad t_u \equiv \min\left\{1, \frac{c_1}{1-\alpha_1}, \frac{c_2}{1-\alpha_2}\right\}.$$

Thus $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} \neq \emptyset$ if and only if $t_l \leq t_u$. It follows directly that for fixed c_1, c_2 the set of pairs $\langle \alpha_1, \alpha_2 \rangle \in [0,1]^2$ for which $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$ is not empty may be characterized thus:

$$\begin{aligned} \mathcal{R}_{c_1,c_2} &\equiv \left\{ \langle \alpha_1, \alpha_2 \rangle \,\middle|\, \mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} \neq \emptyset \right\} \\ &= [0,1]^2 \cap \bigcap_{\substack{i \in \{1,2\} \\ i^* = 3-i}} \{ \langle \alpha_1, \alpha_2 \rangle \mid (\alpha_i - c_i)(\alpha_{i^*} - (1-c_{i^*})) \leq c_i^*(1-c_i) \}. \end{aligned} \tag{5}$$

In fact, as shown in Figure 6 at most one constraint is active, so simplification is possible: let $k = \arg\max_j c_j$, and $k^* = 3-k$, then

$$\mathcal{R}_{c_1,c_2} = [0,1]^2 \cap \{ \langle \alpha_1, \alpha_2 \rangle \mid (\alpha_k - c_k)(\alpha_{k^*} - (1-c_{k^*})) \leq c_k^*(1-c_k) \}.$$

(If $c_1 = c_2$ then $\mathcal{R}_{c_1,c_2} = [0,1]^2$.)

In the two dimensional analysis in §2.2 we observed that for fixed c, as α varied, the line segment would always remain inside two rectangles, as shown in Figure 3. In the three dimensional situation, the line segment (4) will stay within three boxes:

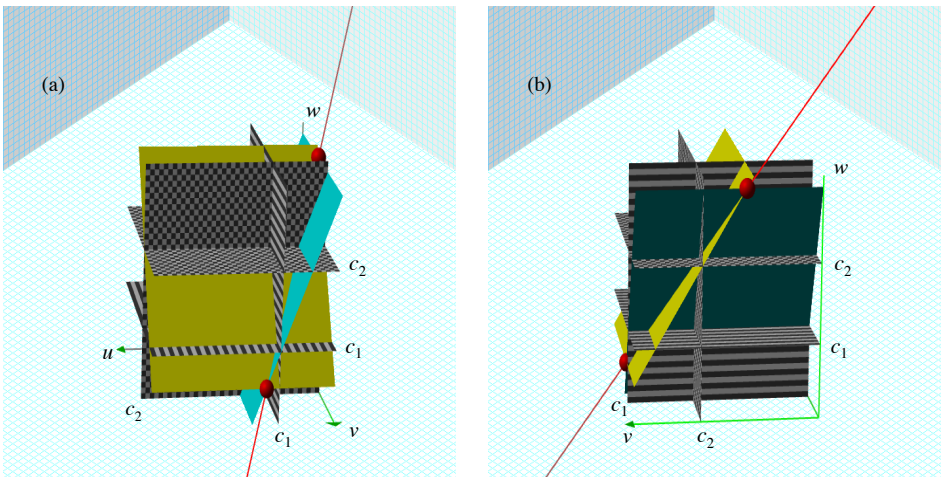

Figure 5: $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$ corresponds to the section of the line between the two marked points; (a) view towards u-w plane; (b) view from v-w plane. (Here $c_1 < c_2$.)

(i) If $c_1 < c_2$ then the line segment (4) is within:

$$([0, c_1] \times [0, c_2] \times [c_2, 1]) \cup ([0, c_1] \times [c_2, 1] \times [c_1, c_2]) \cup ([c_1, 1] \times [c_2, 1] \times [0, c_1]).$$

This may be seen as a 'staircase' with a 'corner' consisting of three blocks, descending clockwise from $\langle 0, 0, 1\rangle$ to $\langle 1, 1, 0\rangle$; see Figure 7(a). The first and second boxes intersect in the line segment joining the points $\langle 0, c_2, c_2\rangle$ and $\langle c_1, c_2, c_2\rangle$; the second and third intersect in the line segment joining $\langle c_1, c_2, c_1\rangle$ and $\langle c_1, 1, c_1\rangle$.

(ii) If $c_1 > c_2$ then the line segment is within:

$$([0, c_1] \times [0, c_2] \times [c_1, 1]) \cup ([c_1, 1] \times [0, c_2] \times [c_2, c_1]) \cup ([c_1, 1] \times [c_2, 1] \times [0, c_2]).$$

This is a 'staircase' of three blocks, descending counter-clockwise from $\langle 0, 0, 1\rangle$ to $\langle 1, 1, 0\rangle$; see Figure 7(b). The first and second boxes intersect in the line segment joining the points $\langle c_1, 0, c_1\rangle$ and $\langle c_1, c_2, c_1\rangle$; the second and third intersect in the line segment joining $\langle c_1, c_2, c_2\rangle$ and $\langle 1, c_2, c_2\rangle$.

(iii) If $c_1 = c_2 = c$ then the 'middle' box disappears and we are left with

$$([0, c] \times [0, c] \times [c, 1]) \cup ([c, 1] \times [c, 1] \times [0, c]).$$

In this case the two boxes touch at the point $\langle c, c, c\rangle$.

Note however, that the number of 'boxes' within which the line segment (4) lies may be 1, 2 or 3 (or 0 if $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)} = \emptyset$). This is in contrast to the simpler case considered in §2.2 where the line segment $\mathcal{Q}_{c,\alpha}$ always intersected exactly two rectangles; see Figure 3.

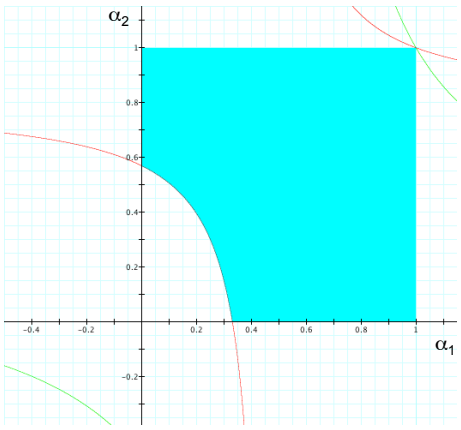

Figure 6: $\mathcal{R}_{c_1,c_2}$ corresponds to the shaded region. The hyperbola of which one arm forms a boundary of this region corresponds to the active constraint; the other hyperbola to the inactive constraint.

3 Characterization of compatible distributions of type

Returning to the Instrumental Variable model introduced in §2, for a given patient the values taken by Y and X are deterministic functions of Z, t_X and t_Y. Consequently, under randomization (2), a distribution over $\mathbb{D}$ determines the conditional distributions $p(x, y \mid z)$ for $z \in \{0, 1\}$. However, since distributions on $\mathbb{D}$ form a 15 dimensional simplex, while $p(x, y \mid z)$ is of dimension 6, it is clear that the reverse does not hold; thus many different distributions over $\mathbb{D}$ give rise to the same distributions $p(x, y \mid z)$. In what follows we precisely characterize the set of distributions over $\mathbb{D}$ corresponding to a given distribution $p(x, y \mid z)$.

We will accomplish this in the following steps:

1. We first characterize the set of distributions π_X on $\mathbb{D}_X$ compatible with a given distribution $p(x \mid z)$.

2. Next we use the technique used for Step 1 to reduce the problem of characterizing distributions $\pi_{Y|X}$ compatible with $p(x, y \mid z)$ to that of characterizing the values of $p(y_x = 1 \mid \mathbf{t}_X)$ compatible with $p(x, y \mid z)$.

3. For a fixed marginal distribution π_X on $\mathbb{D}_X$ we then describe the set of values for $p(y_x = 1 \mid x, \mathbf{t}_X)$ compatible with the observed distribution $p(y \mid x, z)$.

4. In general, some distributions π_X on $\mathbb{D}_X$ and observed distributions $p(y \mid x, z)$ may be incompatible in that there are no compatible values for $p(y_x = 1 \mid \mathbf{t}_X)$. We use this to find the set of distributions π_X on $\mathbb{D}_X$ compatible with $p(y, x \mid z)$ (by restricting the set of distributions found at step 1).

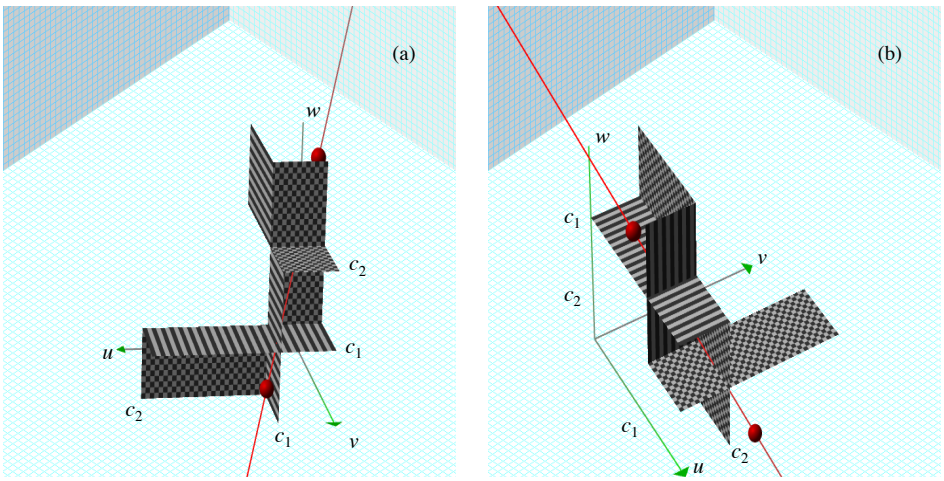

Figure 7: 'Staircases' of three boxes illustrating the possible support for $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$; (a) $c_1 < c_2$; (b) $c_2 < c_1$. Sides of the boxes that are formed by (subsets of) faces of the unit cube are not shown. The line segments shown are illustrative; in general they may not intersect all 3 boxes.

5. Finally we describe the values for $p(y_x = 1 \mid \mathbf{t}_X)$ compatible with the distributions π over $\mathbb{D}_X$ found at the previous step.

We now proceed with the analysis.

3.1 Distributions π_X on $\mathbb{D}_X$ compatible with $p(x \mid z)$

Under random assignment we have

$$\begin{aligned} p(x=1 \mid z=0) &= p(X_{z=0}=1, X_{z=1}=0) + p(X_{z=0}=1, X_{z=1}=1) \\ &= p(\mathrm{DE}) + p(\mathrm{AT}), \end{aligned}$$

$$\begin{aligned} p(x=1 \mid z=1) &= p(X_{z=0}=0, X_{z=1}=1) + p(X_{z=0}=1, X_{z=1}=1) \\ &= p(\mathrm{CO}) + p(\mathrm{AT}). \end{aligned}$$

Letting $U_{i+1} = X_{z=i}$, $i = 0, 1$ and $c_{j+1} = p(x = 1 \mid z = j)$, $j = 0, 1$, it follows directly from the analysis in §2.1 that the set of distributions π_X on $\mathbb{D}_X$ that are compatible with $p(x \mid z)$ are thus given by

$$\mathcal{P}_{c_1,c_2} = \left\{ \begin{array}{lcll} \pi_{\mathrm{AT}} &=& t, & \\ \pi_{\mathrm{DE}} &=& c_1 - t, & \\ \pi_{\mathrm{CO}} &=& c_2 - t, & t \in \left[\max\left\{0, (c_1+c_2)-1\right\}, \min\left\{c_1, c_2\right\}\right] \\ \pi_{\mathrm{NT}} &=& 1 - c_1 - c_2 + t, & \end{array} \right\}. \tag{6}$$

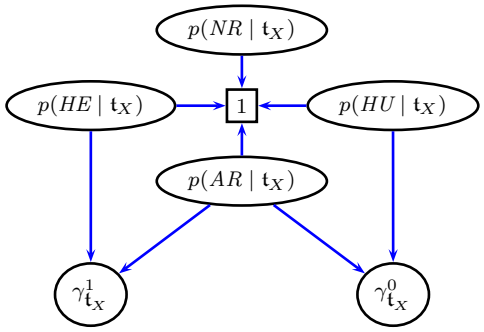

Figure 8: A graph representing the functional dependencies used in the reduction step in §3.2. The rectangular node indicates that the probabilities are required to sum to 1.

3.2 Reduction step in characterizing distributions $\pi_{Y|X}$ compatible with $p(x, y \mid z)$

Suppose that we were able to ascertain the set of possible values for the eight quantities:

$$\gamma^i_{\mathfrak{t}_X} \equiv p(y_{x=i} = 1 \mid \mathfrak{t}_X), \text{ for } i \in \{0, 1\} \text{ and } \mathfrak{t}_X \in \mathbb{D}_X,$$

that are compatible with $p(x, y \mid z)$. Note that $p(y_{x=i} = 1 \mid \mathfrak{t}_X)$ is written as $p(y = 1 \mid \text{do}(x = i), \mathfrak{t}_X)$ using Pearl's do$(\cdot)$ notation. It is then clear that the set of possible distributions $\pi_{Y|X}$ that are compatible with $p(x, y \mid z)$ simply follows from the analysis in §2.1, since

$$\begin{aligned} \gamma^0_{\mathfrak{t}_X} &= p(y_{x=0} = 1 \mid \mathfrak{t}_X) \\ &= p(HU \mid \mathfrak{t}_X) + p(AR \mid \mathfrak{t}_X), \end{aligned}$$

$$\begin{aligned} \gamma^1_{\mathfrak{t}_X} &= p(y_{x=1} = 1 \mid \mathfrak{t}_X) \\ &= p(HE \mid \mathfrak{t}_X) + p(AR \mid \mathfrak{t}_X). \end{aligned}$$

These relationships are also displayed graphically in Figure 8: in this particular graph all children are simple sums of their parents; the boxed 1 represents the 'sum to 1' constraint.

Thus, by §2.1, for given values of $\gamma^i_{\mathfrak{t}_X}$ the set of distributions $\pi_{Y|X}$ is given by:

$$\left\{ \begin{array}{lcl} p(AR \mid \mathfrak{t}_X) & \in & \left[\max\left\{0, (\gamma^0_{\mathfrak{t}_X} + \gamma^1_{\mathfrak{t}_X}) - 1\right\}, \min\left\{\gamma^0_{\mathfrak{t}_X}, \gamma^1_{\mathfrak{t}_X}\right\}\right], \\ p(NR \mid \mathfrak{t}_X) & = & 1 - \gamma^0_{\mathfrak{t}_X} - \gamma^1_{\mathfrak{t}_X} + p(AR \mid \mathfrak{t}_X), \\ p(HE \mid \mathfrak{t}_X) & = & \gamma^1_{\mathfrak{t}_X} - p(AR \mid \mathfrak{t}_X), \\ p(HU \mid \mathfrak{t}_X) & = & \gamma^0_{\mathfrak{t}_X} - p(AR \mid \mathfrak{t}_X) \end{array} \right\}. \tag{7}$$

It follows from the discussion at the end of §2.1 that the values of $\gamma^0_{\mathfrak{t}_X}$ and $\gamma^1_{\mathfrak{t}_X}$ are not restricted by the requirement that there exists a distribution $p(\cdot \mid \mathfrak{t}_X)$ on $\mathbb{D}_Y$. Consequently

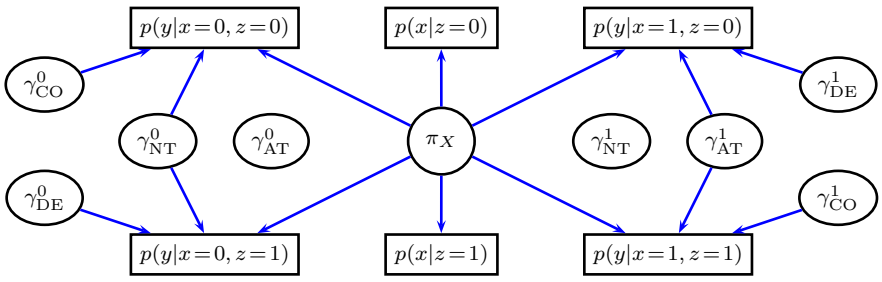

Figure 9: A graph representing the functional dependencies in the analysis of the binary IV model. Rectangular nodes are observed; oval nodes are unknown parameters. See text for further explanation.

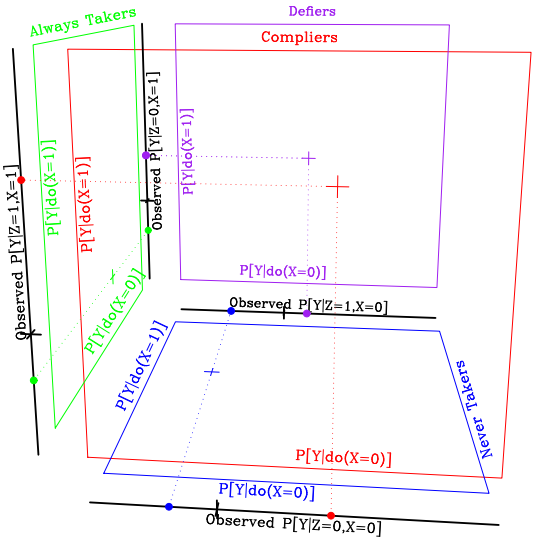

Figure 10: Geometric picture illustrating the relation between the $\gamma^i_{\mathbf{t}_X}$ parameters and $p(y \mid x, z)$. See also Figure 9.

we may proceed in two steps: first we derive the set of values for the eight parameters $\{\gamma^i_{\mathbf{t}_X}\}$ and the distribution on π_X (jointly) without consideration of the parameters for $\pi_{Y|X}$; second we then derive the parameters $\pi_{Y|X}$, as described above.

Finally we note that many causal quantities of interest, such as the average causal effect (ACE), and relative risk (RR) of X on Y, for a given response type $\mathbf{t}_X$, may be expressed in terms of the $\gamma^i_{\mathbf{t}_X}$ parameters:

$$\mathrm{ACE}(\mathbf{t}_X) = \gamma^1_{\mathbf{t}_X} - \gamma^0_{\mathbf{t}_X}, \qquad \mathrm{RR}(\mathbf{t}_X) = \gamma^1_{\mathbf{t}_X} / \gamma^0_{\mathbf{t}_X}.$$

Consequently, for many purposes it may be unnecessary to consider the parameters $\pi_{Y|X}$ at all.

3.3 Values for $\{\gamma^i_{\mathbf{t}_X}\}$ compatible with π_X and $p(y \mid x, z)$

We will call a specification of values for π_X, *feasible for the observed distribution* if (a) π_X lies within the set described in §3.1 of distributions compatible with $p(x \mid z)$ and (b) there exists a set of values for $\gamma^i_{\mathbf{t}_X}$ which results in the distribution $p(y \mid x, z)$.

In the next section we give an explicit characterization of the set of feasible distributions π_X; in this section we characterize the set of values of $\gamma^i_{\mathbf{t}_X}$ compatible with a fixed feasible distribution π_X and $p(y \mid x, z)$.

Proposition 1 *The following equations relate* π_X, γ^0_{CO}, γ^0_{DE}, γ^0_{NT} *to* $p(y \mid x{=}0, z)$*:*

$$p(y{=}1 \mid x{=}0, z{=}0) = (\gamma^0_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}}), \tag{8}$$

$$p(y{=}1 \mid x{=}0, z{=}1) = (\gamma^0_{\mathrm{DE}}\pi_{\mathrm{DE}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{DE}} + \pi_{\mathrm{NT}}), \tag{9}$$

Similarly, the following relate π_X, γ^1_{CO}, γ^1_{DE}, γ^1_{AT} *to* $p(y \mid x{=}1, z)$*:*

$$p(y{=}1 \mid x{=}1, z{=}0) = (\gamma^1_{\mathrm{DE}}\pi_{\mathrm{DE}} + \gamma^1_{\mathrm{AT}}\pi_{\mathrm{AT}})/(\pi_{\mathrm{DE}} + \pi_{\mathrm{AT}}), \tag{10}$$

$$p(y{=}1 \mid x{=}1, z{=}1) = (\gamma^1_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^1_{\mathrm{AT}}\pi_{\mathrm{AT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{AT}}). \tag{11}$$

Equations (8)–(11) are represented in Figure 9. Note that the parameters γ^0_{AT} and γ^1_{NT} are completely unconstrained by the observed distribution since they describe, respectively, the effect of non-exposure ($X = 0$) on Always Takers, and exposure ($X = 1$) on Never Takers, neither of which ever occur. Consequently, the set of possible values for each of these parameters is always $[0, 1]$. Graphically this corresponds to the disconnection of γ^0_{AT} and γ^1_{NT} from the remainder of the graph.

As shown in Proposition 1 the remaining six parameters may be divided into two groups, $\{\gamma^0_{\mathrm{NT}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{CO}}\}$ and $\{\gamma^1_{\mathrm{AT}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{CO}}\}$, depending on whether they relate to unexposed subjects, or exposed subjects. Furthermore, as the graph indicates, for a fixed feasible value of π_X, compatible with the observed distribution $p(x, y \mid z)$ (assuming such exists), these two sets are variation independent. Thus, for a fixed feasible value of π_X we may analyze each of these sets separately.

A geometric picture of equations (8)–(11) is given in Figure 10: there is one square for each compliance type, with axes corresponding to $\gamma^0_{\mathbf{t}_X}$ and $\gamma^1_{\mathbf{t}_X}$; the specific value of $\langle \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X} \rangle$ is given by a cross in the square. There are four lines corresponding to the four observed quantities $p(y = 1 \mid x, z)$. Each of these observed quantities, which is denoted by a cross on the respective line, is a weighted average of two $\gamma^i_{\mathbf{t}_X}$ parameters, with weights given by π_X (the weights are not depicted explicitly).

Proof of Proposition 1: We prove (8); the other proofs are similar. Subjects for whom $X = 0$ and $Z = 0$ are either Never Takers or Compliers. Hence

$$\begin{aligned}
p(y=1 \mid x=0, z=0) &= p(y=1 \mid x=0, z=0, \mathfrak{t}_X=\mathrm{NT})p(\mathfrak{t}_X=\mathrm{NT} \mid x=0, z=0) \\
&\quad + p(y=1 \mid x=0, z=0, \mathfrak{t}_X=\mathrm{CO})p(\mathfrak{t}_X=\mathrm{CO} \mid x=0, z=0) \\
&= p(y_{x=0}=1 \mid x=0, z=0, \mathfrak{t}_X=\mathrm{NT})p(\mathfrak{t}_X=\mathrm{NT} \mid \mathfrak{t}_X \in \{\mathrm{CO}, \mathrm{NT}\}) \\
&\quad + p(y_{x=0}=1 \mid x=0, z=0, \mathfrak{t}_X=\mathrm{CO})p(\mathfrak{t}_X=\mathrm{CO} \mid \mathfrak{t}_X \in \{\mathrm{CO}, \mathrm{NT}\}) \\
&= p(y_{x=0}=1 \mid z=0, \mathfrak{t}_X=\mathrm{NT}) \times \pi_{\mathrm{NT}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\
&\quad + p(y_{x=0}=1 \mid z=0, \mathfrak{t}_X=\mathrm{CO}) \times \pi_{\mathrm{CO}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\
&= p(y_{x=0}=1 \mid \mathfrak{t}_X=\mathrm{NT}) \times \pi_{\mathrm{NT}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\
&\quad + p(y_{x=0}=1 \mid \mathfrak{t}_X=\mathrm{CO}) \times \pi_{\mathrm{CO}}/(\pi_{\mathrm{NT}} + \pi_{\mathrm{CO}}) \\
&= (\gamma^0_{\mathrm{CO}}\pi_{\mathrm{CO}} + \gamma^0_{\mathrm{NT}}\pi_{\mathrm{NT}})/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}}).
\end{aligned}$$

Here the first equality is by the chain rule of probability; the second follows by consistency; the third follows since Compliers and Never Takers have $X = 0$ when $Z = 0$; the fourth follows by randomization (2). □

Values for $\gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}}$ compatible with a feasible π_X

Since (8) and (9) correspond to three quantities with two averages specified, we may apply the analysis in §2.3, taking $\alpha_1 = \pi_{\mathrm{CO}}/(\pi_{\mathrm{CO}} + \pi_{\mathrm{NT}})$, $\alpha_2 = \pi_{\mathrm{DE}}/(\pi_{\mathrm{DE}} + \pi_{\mathrm{NT}})$, $c_i = p(y=1 \mid x=0, z=i-1)$ for $i = 1, 2$, $u = \gamma^0_{\mathrm{CO}}$, $v = \gamma^0_{\mathrm{DE}}$ and $w = \gamma^0_{\mathrm{NT}}$. Under this substitution, the set of possible values for $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ is then given by $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$.

Values for $\gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}}$ compatible with a feasible π_X

Likewise since (10) and (11) contain three quantities with two averages specified we again apply the analysis from §2.3, taking $\alpha_1 = \pi_{\mathrm{CO}}/(\pi_{\mathrm{CO}} + \pi_{\mathrm{AT}})$, $\alpha_2 = \pi_{\mathrm{DE}}/(\pi_{\mathrm{DE}} + \pi_{\mathrm{AT}})$, $c_i = p(y=1 \mid x=1, z=2-i)$ for $i = 1, 2$, $u = \gamma^1_{\mathrm{CO}}$, $v = \gamma^1_{\mathrm{DE}}$ and $w = \gamma^1_{\mathrm{AT}}$. The set of possible values for $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ is then given by $\mathcal{Q}_{(c_1,\alpha_1)(c_2,\alpha_2)}$.

3.4 Values of π_X compatible with $p(x, y \mid z)$

In §3.1 we characterized the distributions π_X compatible with $p(x \mid z)$ as a one dimensional subspace of the three dimensional simplex, parameterized in terms of $t \equiv \pi_{\mathrm{AT}}$; see (6). We now incorporate the additional constraints on π_X that arise from $p(y \mid x, z)$. These occur

because some distributions π_X, though compatible with $p(x \mid z)$, lead to an empty set of values for $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT} \rangle$ or $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ and thus are infeasible.

3.4.1 Constraints on π_X arising from $p(y \mid x = 0, z)$

Building on the analysis in §3.3 the set of values for

$$\begin{aligned} \langle \alpha_1, \alpha_2 \rangle &= \langle \pi_{\rm CO}/(\pi_{\rm CO} + \pi_{\rm NT}), \pi_{\rm DE}/(\pi_{\rm DE} + \pi_{\rm NT}) \rangle \\ &= \langle \pi_{\rm CO}/p_{x_0|z_0}, \pi_{\rm DE}/p_{x_0|z_0} \rangle \end{aligned} \tag{12}$$

compatible with $p(y \mid x=0, z)$ (i.e. for which the corresponding set of values for $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ is non-empty) is given by $\mathcal{R}_{c^*_1, c^*_2}$, where $c^*_i = p(y=1 \mid x=0, z=i-1)$, $i = 1, 2$ (see §2.3). The inequalities defining $\mathcal{R}_{c^*_1, c^*_2}$ may be translated into upper bounds on $t \equiv \pi_{\rm AT}$ in (6), as follows:

$$t \leq \min \left\{ 1 - \sum_{j \in \{0,1\}} p(y=j, x=0 \mid z=j),\ 1 - \sum_{k \in \{0,1\}} p(y=k, x=0 \mid z=1-k) \right\}. \tag{13}$$

Proof: The analysis in §3.3 implied that for $\mathcal{R}_{c^*_1, c^*_2} \neq \emptyset$ we require

$$\frac{c^*_1 - \alpha_1}{1 - \alpha_1} \leq \frac{c^*_2}{1 - \alpha_2} \quad \text{and} \quad \frac{c^*_2 - \alpha_2}{1 - \alpha_2} \leq \frac{c^*_1}{1 - \alpha_1}. \tag{14}$$

Taking the first of these and plugging in the definitions of c^*_1, c^*_2, α_1 and α_2 from (12) gives:

$$\begin{aligned} & \frac{p_{y_1|x_0,z_0} - (\pi_{\rm CO}/p_{x_0|z_0})}{1 - (\pi_{\rm CO}/p_{x_0|z_0})} \leq \frac{p_{y_1|x_0,z_1}}{1 - (\pi_{\rm DE}/p_{x_0|z_1})} \\ (\Leftrightarrow) \quad & (p_{y_1|x_0,z_0} - (\pi_{\rm CO}/p_{x_0|z_0}))(1 - (\pi_{\rm DE}/p_{x_0|z_1})) \leq p_{y_1|x_0,z_1}(1 - (\pi_{\rm CO}/p_{x_0|z_0})) \\ (\Leftrightarrow) \quad & (p_{y_1,x_0|z_0} - \pi_{\rm CO})(p_{x_0|z_1} - \pi_{\rm DE}) \leq p_{y_1,x_0|z_1}(p_{x_0|z_0} - \pi_{\rm CO}). \end{aligned}$$

But $p_{x_0|z_1} - \pi_{\rm DE} = p_{x_0|z_0} - \pi_{\rm CO} = \pi_{\rm NT}$, hence these terms may be cancelled to give:

$$\begin{aligned} & (p_{y_1,x_0|z_0} - \pi_{\rm CO}) \leq p_{y_1,x_0|z_1} \\ (\Leftrightarrow) \quad & \pi_{\rm AT} - p_{x_1|z_1} \leq p_{y_1,x_0|z_1} - p_{y_1,x_0|z_0} \\ (\Leftrightarrow) \quad & \pi_{\rm AT} \leq 1 - p_{y_0,x_0|z_1} - p_{y_1,x_0|z_0}. \end{aligned}$$

A similar argument applied to the second constraint in (14) to derive that

$$\pi_{\rm AT} \leq 1 - p_{y_0,x_0|z_0} - p_{y_1,x_0|z_1},$$

as required. □

3.4.2 Constraints on π_X arising from $p(y \mid x = 1, z)$

Similarly using the analysis in §3.3 the set of values for

$$\langle \alpha_1, \alpha_2 \rangle = \langle \pi_{\rm CO}/(\pi_{\rm CO} + \pi_{\rm AT}), \pi_{\rm DE}/(\pi_{\rm DE} + \pi_{\rm AT}) \rangle$$

compatible with $p(y \mid x=1, z)$ (i.e. that the corresponding set of values for $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT} \rangle$ is non-empty) is given by $\mathcal{R}_{c_1^{**}, c_2^{**}}$, where $c_i^{**} = p(y = 1 \mid x = 1, z = 2 - i)$, $i = 1, 2$ (see §2.3). Again, we translate the inequalities which define $\mathcal{R}_{c_1^{**}, c_2^{**}}$ into further upper bounds on $t = \pi_{\rm AT}$ in (6):

$$t \leq \min \left\{ \sum_{j \in \{0,1\}} p(y{=}j, x{=}1 \mid z{=}j), \sum_{k \in \{0,1\}} p(y{=}k, x{=}1 \mid z{=}1{-}k) \right\}. \tag{15}$$

The proof that these inequalities are implied, is very similar to the derivation of the upper bounds on $\pi_{\rm AT}$ arising from $p(y \mid x = 0, z)$ considered above.

3.4.3 The distributions π_X compatible with the observed distribution

It follows that the set of distributions on $\mathbb{D}_X$ that are compatible with the observed distribution, which we denote $\mathcal{P}_X$, may be given thus:

$$\mathcal{P}_X = \left\{ \begin{array}{rcl} \pi_{\rm AT} & \in & [l\pi_{\rm AT}, u\pi_{\rm AT}], \\ \pi_{\rm NT}(\pi_{\rm AT}) & = & 1 - p(x = 1 \mid z = 0) - p(x = 1 \mid z = 1) + \pi_{\rm AT}, \\ \pi_{\rm CO}(\pi_{\rm AT}) & = & p(x = 1 \mid z = 1) - \pi_{\rm AT}, \\ \pi_{\rm DE}(\pi_{\rm AT}) & = & p(x = 1 \mid z = 0) - \pi_{\rm AT} \end{array} \right\}, \tag{16}$$

where

$$l\pi_{\rm AT} = \max\{0, p(x = 1 \mid z = 0) + p(x = 1 \mid z = 1) - 1\};$$

$$u\pi_{\rm AT} = \min \left\{ \begin{array}{ll} p(x = 1 \mid z = 0), & p(x = 1 \mid z = 1), \\ 1 - \sum_j p(y{=}j, x{=}0 \mid z{=}j), & 1 - \sum_k p(y{=}k, x{=}0 \mid z{=}1{-}k), \\ \sum_j p(y{=}j, x{=}1 \mid z{=}j), & \sum_k p(y{=}k, x{=}1 \mid z{=}1{-}k) \end{array} \right\}.$$

Observe that unlike the upper bound, the lower bound on $\pi_{\rm AT}$ (and $\pi_{\rm NT}$) obtained from $p(x, y \mid z)$ is the same as the lower bound derived from $p(x \mid z)$ alone.
We define $\pi_X(\pi_{\rm AT}) \equiv \langle \pi_{\rm NT}(\pi_{\rm AT}), \pi_{\rm CO}(\pi_{\rm AT}), \pi_{\rm DE}(\pi_{\rm AT}), \pi_{\rm AT} \rangle$, for use below. Note the following:

Proposition 2 *When $\pi_{\rm AT}$ (equivalently $\pi_{\rm NT}$) is minimized then either $\pi_{\rm NT} = 0$ or $\pi_{\rm AT} = 0$.*

Proof: This follows because, by the expression for $l\pi_{\rm AT}$, either $l\pi_{\rm AT} = 0$, or $l\pi_{\rm AT} = p(x = 1 \mid z = 0) + p(x = 1 \mid z = 1) - 1$, in which case $l\pi_{\rm NT} = 0$ by (16). □

4 Projections

The analysis in §3 provides a complete description of the set of distributions over $\mathbb{D}$ compatible with a given observed distribution. In particular, equation (16) describes the one dimensional set of compatible distributions over $\mathbb{D}_X$; in §3.3 we first gave a description of the one dimensional set of values over $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ compatible with the observed distribution and a specific feasible distribution π_X over $\mathbb{D}_X$; we then described the one dimensional set of values for $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT} \rangle$. Varying π_X over the set $\mathcal{P}_X$ of feasible distributions over $\mathbb{D}_X$, describes a set of lines, forming two two-dimensional manifolds which represent the space of possible values for $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ and likewise for $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT} \rangle$. As noted previously, the parameters $\gamma^0_{\rm AT}$ and $\gamma^1_{\rm NT}$ are unconstrained by the observed data. Finally, if there is interest in distributions over response types, there is a one-dimensional set of such distributions associated with each possible pair of values from $\gamma^0_{\mathbf{t}_X}$ and $\gamma^1_{\mathbf{t}_X}$.

For the purposes of visualization it is useful to look at projections. There are many such projections that could be considered, here we focus on projections that display the relation between the possible values for π_X and $\gamma^x_{\mathbf{t}_X}$. See Figure 11.

We make the following definition:

$$\alpha^{ij}_{\mathbf{t}_X}(\pi_X) \;\equiv\; p(\mathbf{t}_X \mid X_{z=i} = j),$$

where $\pi_X = \langle \pi_{\rm NT}, \pi_{\rm CO}, \pi_{\rm DE}, \pi_{\rm AT} \rangle \in \mathcal{P}_X$, as before. For example, $\alpha^{00}_{\rm NT}(\pi_X) = \pi_{\rm NT}/(\pi_{\rm NT} + \pi_{\rm CO})$, $\alpha^{10}_{\rm NT}(\pi_X) = \pi_{\rm NT}/(\pi_{\rm NT} + \pi_{\rm DE})$.

4.1 Upper and Lower bounds on $\gamma^x_{\mathbf{t}_X}$ as a function of π_X

We use the following notation to refer to the upper and lower bounds on $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ that were derived earlier. If π_X is such that $\pi_{\rm NT} > 0$, so $\alpha^{00}_{\rm NT}, \alpha^{10}_{\rm NT} > 0$ then we define:

$$\begin{aligned} l\gamma^0_{\rm NT}(\pi_X) &\equiv \max\left\{0, \frac{p_{y_1|x_0z_0} - \alpha^{00}_{\rm CO}(\pi_X)}{\alpha^{00}_{\rm NT}(\pi_X)}, \frac{p_{y_1|x_0z_1} - \alpha^{10}_{\rm DE}(\pi_X)}{\alpha^{10}_{\rm NT}(\pi_X)}\right\}, \\ u\gamma^0_{\rm NT}(\pi_X) &\equiv \min\left\{\frac{p_{y_1|x_0z_0}}{\alpha^{00}_{\rm NT}(\pi_X)}, \frac{p_{y_1|x_0z_1}}{\alpha^{10}_{\rm NT}(\pi_X)}, 1\right\}, \end{aligned}$$

while if $\pi_{\rm NT} = 0$ then we define $l\gamma^0_{\rm NT}(\pi_X) \equiv 0$ and $u\gamma^0_{\rm NT}(\pi_X) \equiv 1$. Similarly, if π_X is such that $\pi_{\rm AT} > 0$ then we define:

$$\begin{aligned} l\gamma^1_{\rm AT}(\pi_X) &\equiv \max\left\{0, \frac{p_{y_1|x_1z_1} - \alpha^{11}_{\rm CO}(\pi_X)}{\alpha^{11}_{\rm AT}(\pi_X)}, \frac{p_{y_1|x_1z_0} - \alpha^{01}_{\rm DE}(\pi_X)}{\alpha^{01}_{\rm AT}(\pi_X)}\right\}, \\ u\gamma^1_{\rm AT}(\pi_X) &\equiv \min\left\{\frac{p_{y_1|x_1z_1}}{\alpha^{11}_{\rm AT}(\pi_X)}, \frac{p_{y_1|x_1z_0}}{\alpha^{01}_{\rm AT}(\pi_X)}, 1\right\}, \end{aligned}$$

	Lower Bound	Upper Bound
$\gamma^0_{\rm NT}$	$l\gamma^0_{\rm NT}(\pi_X)$	$u\gamma^0_{\rm NT}(\pi_X)$
$\gamma^0_{\rm CO}$	$(p_{y_1\|x_0z_0} - u\gamma^0_{\rm NT}(\pi_X)\cdot\alpha^{00}_{\rm NT})/\alpha^{00}_{\rm CO}$	$(p_{y_1\|x_0z_0} - l\gamma^0_{\rm NT}(\pi_X)\cdot\alpha^{00}_{\rm NT})/\alpha^{00}_{\rm CO}$
$\gamma^0_{\rm DE}$	$(p_{y_1\|x_0z_1} - u\gamma^0_{\rm NT}(\pi_X)\cdot\alpha^{10}_{\rm NT})/\alpha^{10}_{\rm DE}$	$(p_{y_1\|x_0z_1} - l\gamma^0_{\rm NT}(\pi_X)\cdot\alpha^{10}_{\rm NT})/\alpha^{10}_{\rm DE}$
$\gamma^0_{\rm AT}$	0	1
$\gamma^1_{\rm NT}$	0	1
$\gamma^1_{\rm CO}$	$(p_{y_1\|x_1z_1} - u\gamma^1_{\rm AT}(\pi_X)\cdot\alpha^{11}_{\rm AT})/\alpha^{11}_{\rm CO}$	$(p_{y_1\|x_1z_1} - l\gamma^1_{\rm AT}(\pi_X)\cdot\alpha^{11}_{\rm AT})/\alpha^{11}_{\rm CO}$
$\gamma^1_{\rm DE}$	$(p_{y_1\|x_1z_0} - u\gamma^1_{\rm AT}(\pi_X)\cdot\alpha^{01}_{\rm AT})/\alpha^{01}_{\rm DE}$	$(p_{y_1\|x_1z_0} - l\gamma^1_{\rm AT}(\pi_X)\cdot\alpha^{01}_{\rm AT})/\alpha^{01}_{\rm DE}$
$\gamma^1_{\rm AT}$	$l\gamma^1_{\rm AT}(\pi_X)$	$u\gamma^1_{\rm AT}(\pi_X)$

Table 2: Upper and Lower bounds on $\gamma^x_{\mathbf{t}_X}$, as a function of $\pi_X \in \mathcal{P}_X$. If for some π_X an expression giving a lower bound for a quantity is undefined then the lower bound is 0; conversely if an expression for an upper bound is undefined then the upper bound is 1.

while if $\pi_{\rm AT} = 0$ then let $l\gamma^1_{\rm AT}(\pi_X) \equiv 0$ and $u\gamma^1_{\rm AT}(\pi_X) \equiv 1$.

We note that Table 2 summarizes the upper and lower bounds, as a function of $\pi_X \in \mathcal{P}_X$, on each of the eight parameters $\gamma^x_{\mathbf{t}_X}$ that were derived earlier in §3.3. These are shown by the thicker lines on each of the plots forming the upper and lower boundaries in Figure 11 ($\gamma^0_{\rm AT}$ and $\gamma^1_{\rm NT}$ are not shown in the Figure).

The upper and lower bounds on $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ are relatively simple:

Proposition 3 *$l\gamma^0_{\rm NT}(\pi_X)$ and $l\gamma^1_{\rm AT}(\pi_X)$ are non-decreasing in $\pi_{\rm AT}$ and $\pi_{\rm NT}$. Likewise $u\gamma^0_{\rm NT}(\pi_X)$ and $u\gamma^1_{\rm AT}(\pi_X)$ are non-increasing in $\pi_{\rm AT}$ and $\pi_{\rm NT}$.*

Proof: We first consider $l\gamma^0_{\rm NT}$. By (16), $\pi_{\rm NT} = 1 - p(x = 1 \mid z = 0) - p(x = 1 \mid z = 1) + \pi_{\rm AT}$, hence a function is non-increasing [non-decreasing] in $\pi_{\rm AT}$ iff it is non-increasing [non-decreasing] in $\pi_{\rm NT}$. Observe that for $\pi_{\rm NT} > 0$,

$$\begin{aligned}(p_{y_1|x_0z_0} - \alpha^{00}_{\rm CO}(\pi_X))/\alpha^{00}_{\rm NT}(\pi_X) &= \left(p_{y_1|x_0z_0}(\pi_{\rm NT} + \pi_{\rm CO}) - \pi_{\rm CO}\right)/\pi_{\rm NT} \\ &= p_{y_1|x_0z_0} - p_{y_0|x_0z_0}(\pi_{\rm CO}/\pi_{\rm NT}) \\ &= p_{y_1|x_0z_0} + p_{y_0|x_0z_0}(1 - (p_{x_0|z_0}/\pi_{\rm NT}))\end{aligned}$$

which is non-decreasing in $\pi_{\rm NT}$. Similarly,

$$(p_{y_1|x_0z_1} - \alpha^{10}_{\rm DE}(\pi_X))/\alpha^{10}_{\rm NT}(\pi_X) = p_{y_1|x_0z_1} + p_{y_0|x_0z_1}(1 - (p_{x_0|z_1}/\pi_{\rm NT})).$$

The conclusion follows since the maximum of a set of non-decreasing functions is non-decreasing.

The other arguments are similar. □

We note that the bounds on $\gamma^x_{\rm CO}$ and $\gamma^x_{\rm DE}$ need not be monotonic in $\pi_{\rm AT}$.

Proposition 4 *Let $\pi_X^{\min}$ be the distribution in $\mathcal{P}_X$ for which $\pi_{\rm AT}$ and $\pi_{\rm NT}$ are minimized then either:*

(1) $\pi_{\rm NT}^{\min} = 0$, *hence* $l\gamma^0_{\rm NT}(\pi_X^{\min}) = 0$ *and* $u\gamma^0_{\rm NT}(\pi_X^{\min}) = 1$; *or*

(2) $\pi_{\rm AT}^{\min} = 0$, *hence* $l\gamma^1_{\rm AT}(\pi_X^{\min}) = 0$ *and* $u\gamma^1_{\rm AT}(\pi_X^{\min}) = 1$.

Proof: This follows from Proposition 2, and the fact that if $\pi_{\mathbf{t}_X} = 0$ then $\gamma^i_{\mathbf{t}_X}$ is not identified (for any i). □

4.2 Upper and Lower bounds on $p(\mathrm{AT})$ as a function of $\gamma^0_{\rm NT}$

The expressions given in Table 2 allow the range of values for each $\gamma^i_{\mathbf{t}_X}$ to be determined as a function of π_X, giving the upper and lower bounding curves in Figure 11. However it follows directly from (8) and (9) that there is a bijection between the three shapes shown for $\gamma^0_{\rm CO}$, $\gamma^0_{\rm DE}$ and $\gamma^0_{\rm NT}$ (top row of Figure 11). In this section we describe this bijection by deriving curves corresponding to fixed values of $\gamma^0_{\rm NT}$ that are displayed in the plots for $\gamma^0_{\rm CO}$ and $\gamma^0_{\rm DE}$. Similarly it follows from (10) and (11) that there is a bijection between the three shapes shown for $\gamma^1_{\rm CO}$, $\gamma^1_{\rm DE}$, $\gamma^1_{\rm AT}$ (bottom row of Figure 11). Correspondingly we add curves to the plots for $\gamma^1_{\rm CO}$ and $\gamma^1_{\rm DE}$ corresponding to fixed values of $\gamma^1_{\rm AT}$. (The expressions in this section are used solely to add these curves and are not used elsewhere.)

As described earlier, for a given distribution $\pi_X \in \mathcal{P}_X$ the set of values for $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ forms a one dimensional subspace. For a given π_X if $\pi_{\rm CO} > 0$ then $\gamma^0_{\rm CO}$ is a deterministic function of $\gamma^0_{\rm NT}$, likewise for $\gamma^0_{\rm DE}$.

It follows from Proposition 3 that the range of values for $\gamma^0_{\rm NT}$ when $\pi_X = \pi_X^{\min}$ contains the range of possible values for $\gamma^0_{\rm NT}$ for any other $\pi_X \in \mathcal{P}_X$. The same holds for $\gamma^1_{\rm AT}$. Thus for any given possible value of $\gamma^0_{\rm NT}$, the minimum compatible value of $\pi_{\rm AT} = l\pi_{\rm AT} \equiv \max\left\{0, p_{x_1|z_0} + p_{x_1|z_1} - 1\right\}$. This is reflected in the plots in Figure 11 for $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ in that the left hand endpoints of the thinner lines (lying between the upper and lower bounds) all lie on the same vertical line for which $\pi_{\rm AT}$ is minimized.

In contrast the upper bounds on $\pi_{\rm AT}$ vary as a function of $\gamma^0_{\rm NT}$ (also $\gamma^1_{\rm AT}$). The upper bound for $\pi_{\rm AT}$ as a function of $\gamma^0_{\rm NT}$ occurs when one of the thinner horizontal lines in the

plot for $\gamma^0_{\rm NT}$ in Figure 11 intersects either $u\gamma^0_{\rm NT}(\pi_X)$, $l\gamma^0_{\rm NT}(\pi_X)$, or the vertical line given by the global upper bound, $u\pi_{\rm AT}$, on $\pi_{\rm AT}$:

$$\begin{aligned} u\pi_{\rm AT}(\gamma^0_{\rm NT}) &\equiv \max\left\{\pi_{\rm AT} \mid \gamma^0_{\rm NT} \in [l\gamma^0_{\rm NT}(\pi_X), u\gamma^0_{\rm NT}(\pi_X)]\right\} \\ &= \min\left\{ p_{x_1|z_1} - p_{x_0|z_0}\left(1 - \frac{p_{y_1|x_0z_0}}{\gamma^0_{\rm NT}}\right),\ p_{x_1|z_0} - p_{x_0|z_1}\left(1 - \frac{p_{y_1|x_0z_1}}{\gamma^0_{\rm NT}}\right), \right. \\ &\qquad \left. p_{x_1|z_1} - p_{x_0|z_0}\left(1 - \frac{p_{y_0|x_0z_0}}{1-\gamma^0_{\rm NT}}\right),\ p_{x_1|z_0} - p_{x_0|z_1}\left(1 - \frac{p_{y_0|x_0z_1}}{1-\gamma^0_{\rm NT}}\right), u\pi_{\rm AT} \right\}; \end{aligned}$$

similarly we have

$$\begin{aligned} u\pi_{\rm AT}(\gamma^1_{\rm AT}) &\equiv \max\left\{\pi_{\rm AT} \mid \gamma^1_{\rm AT} \in [l\gamma^0_{\rm AT}(\pi_X), u\gamma^1_{\rm AT}(\pi_X)]\right\} \\ &= \min\left\{ u\pi_{\rm AT}, \frac{p_{x_1|z_1}p_{y_1|x_1z_1}}{\gamma^1_{\rm AT}}, \frac{p_{x_1|z_0}p_{y_1|x_1z_0}}{\gamma^1_{\rm AT}}, \frac{p_{x_1|z_1}p_{y_0|x_1z_1}}{1-\gamma^1_{\rm AT}}, \frac{p_{x_1|z_0}p_{y_0|x_1z_0}}{1-\gamma^1_{\rm AT}} \right\}. \end{aligned}$$

The curves added to the unexposed plots for Compliers and Defiers in Figure 11 are as follows:

$$\begin{aligned} \gamma^0_{\rm CO}(\pi_X, \gamma^0_{\rm NT}) &\equiv (p_{y_1|x_0z_0} - \gamma^0_{\rm NT}\cdot\alpha^{00}_{\rm NT})/\alpha^{00}_{\rm CO}, \\ c\gamma^0_{\rm CO}(\pi_{\rm AT}, \gamma^0_{\rm NT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^0_{\rm CO}(\pi_X(\pi_{\rm AT}), \gamma^0_{\rm NT})\rangle\}; \end{aligned} \tag{17}$$

$$\begin{aligned} \gamma^0_{\rm DE}(\pi_X, \gamma^0_{\rm NT}) &\equiv (p_{y_1|x_0z_1} - \gamma^0_{\rm NT}\cdot\alpha^{10}_{\rm NT})/\alpha^{10}_{\rm DE}, \\ c\gamma^0_{\rm DE}(\pi_{\rm AT}, \gamma^0_{\rm NT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^0_{\rm DE}(\pi_X(\pi_{\rm AT}), \gamma^0_{\rm NT})\rangle\}; \end{aligned} \tag{18}$$

for $\gamma^0_{\rm NT} \in [l\gamma^0_{\rm NT}(\pi^{\min}_X), u\gamma^0_{\rm NT}(\pi^{\min}_X)]$; $\pi_{\rm AT} \in [l\pi_{\rm AT}, u\pi_{\rm AT}(\gamma^0_{\rm NT})]$. The curves added to the exposed plots for Compliers and Defiers in Figure 11 are given by:

$$\begin{aligned} \gamma^1_{\rm CO}(\pi_X, \gamma^1_{\rm AT}) &\equiv (p_{y_1|x_1z_1} - \gamma^1_{\rm AT}\cdot\alpha^{11}_{\rm AT})/\alpha^{11}_{\rm CO}, \\ c\gamma^1_{\rm DE}(\pi_{\rm AT}, \gamma^1_{\rm AT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^1_{\rm CO}(\pi_X(\pi_{\rm AT}), \gamma^1_{\rm AT})\rangle\}; \end{aligned} \tag{19}$$

$$\begin{aligned} \gamma^1_{\rm DE}(\pi_X, \gamma^1_{\rm AT}) &\equiv (p_{y_1|x_1z_0} - \gamma^1_{\rm AT}\cdot\alpha^{01}_{\rm AT})/\alpha^{01}_{\rm DE}, \\ c\gamma^1_{\rm DE}(\pi_{\rm AT}, \gamma^1_{\rm AT}) &\equiv \{\langle \pi_{\rm AT}, \gamma^1_{\rm DE}(\pi_X(\pi_{\rm AT}), \gamma^1_{\rm AT})\rangle\}; \end{aligned} \tag{20}$$

for $\gamma^1_{\rm AT} \in [l\gamma^1_{\rm AT}(\pi^{\min}_X), u\gamma^1_{\rm AT}(\pi^{\min}_X)]$; $\pi_{\rm AT} \in [l\pi_{\rm AT}, u\pi_{\rm AT}(\gamma^1_{\rm AT})]$.

4.3 Example: Flu Data

To illustrate some of the constructions described we consider the influenza vaccine dataset [McDonald, Hiu, and Tierney 1992] previously analyzed by [Hirano, Imbens, Rubin, and Zhou 2000]; see Table 3. Here the instrument Z was whether a patient's physician was sent a card asking him to remind patients to obtain flu shots, or not; X is whether or not the

Table 3: Flu Vaccine Data from [McDonald, Hiu, and Tierney 1992].

Z	X	Y	count
0	0	0	99
0	0	1	1027
0	1	0	30
0	1	1	233
1	0	0	84
1	0	1	935
1	1	0	31
1	1	1	422
			2,861

patient did in fact get a flu shot. Finally $Y = 1$ indicates that a patient was *not* hospitalized. Unlike the analysis of [Hirano, Imbens, Rubin, and Zhou 2000] we ignore baseline covariates, and restrict attention to displaying the set of parameters of the IV model that are compatible with the empirical distribution.

The set of values for π_X vs. $\langle \gamma^0_{\rm CO}, \gamma^0_{\rm DE}, \gamma^0_{\rm NT} \rangle$ (upper row), and π_X vs. $\langle \gamma^1_{\rm CO}, \gamma^1_{\rm DE}, \gamma^1_{\rm AT} \rangle$ corresponding to the empirical distribution for $p(x, y \mid z)$ are shown in Figure 11. The empirical distribution is not consistent with there being no Defiers (though the scales in Figure 11 show 0 as one endpoint for the proportion $\pi_{\rm DE}$ this is merely a consequence of the significant digits displayed; in fact the true lower bound on this proportion is 0.0005).

We emphasize that this analysis merely derives the logical consequences of the empirical distribution under the IV model and ignores sampling variability.

5 Bounding Average Causal Effects

We may use the results above to obtain bounds on average causal effects, for different complier strata:

$$\begin{aligned} {\rm ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}) &\equiv \gamma^1_{\mathbf{t}_X}(\pi_X) - \gamma^0_{\mathbf{t}_X}(\pi_X), \\ l{\rm ACE}_{\mathbf{t}_X}(\pi_X) &\equiv \min_{\gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}} {\rm ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}), \\ u{\rm ACE}_{\mathbf{t}_X}(\pi_X) &\equiv \max_{\gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}} {\rm ACE}_{\mathbf{t}_X}(\pi_X, \gamma^0_{\mathbf{t}_X}, \gamma^1_{\mathbf{t}_X}), \end{aligned}$$

as a function of a feasible distribution π_X; see Table 5. As shown in the table, the values of $\gamma^0_{\rm NT}$ and $\gamma^1_{\rm AT}$ which maximize (minimize) ${\rm ACE}_{\rm CO}$ and ${\rm ACE}_{\rm DE}$ are those which minimize

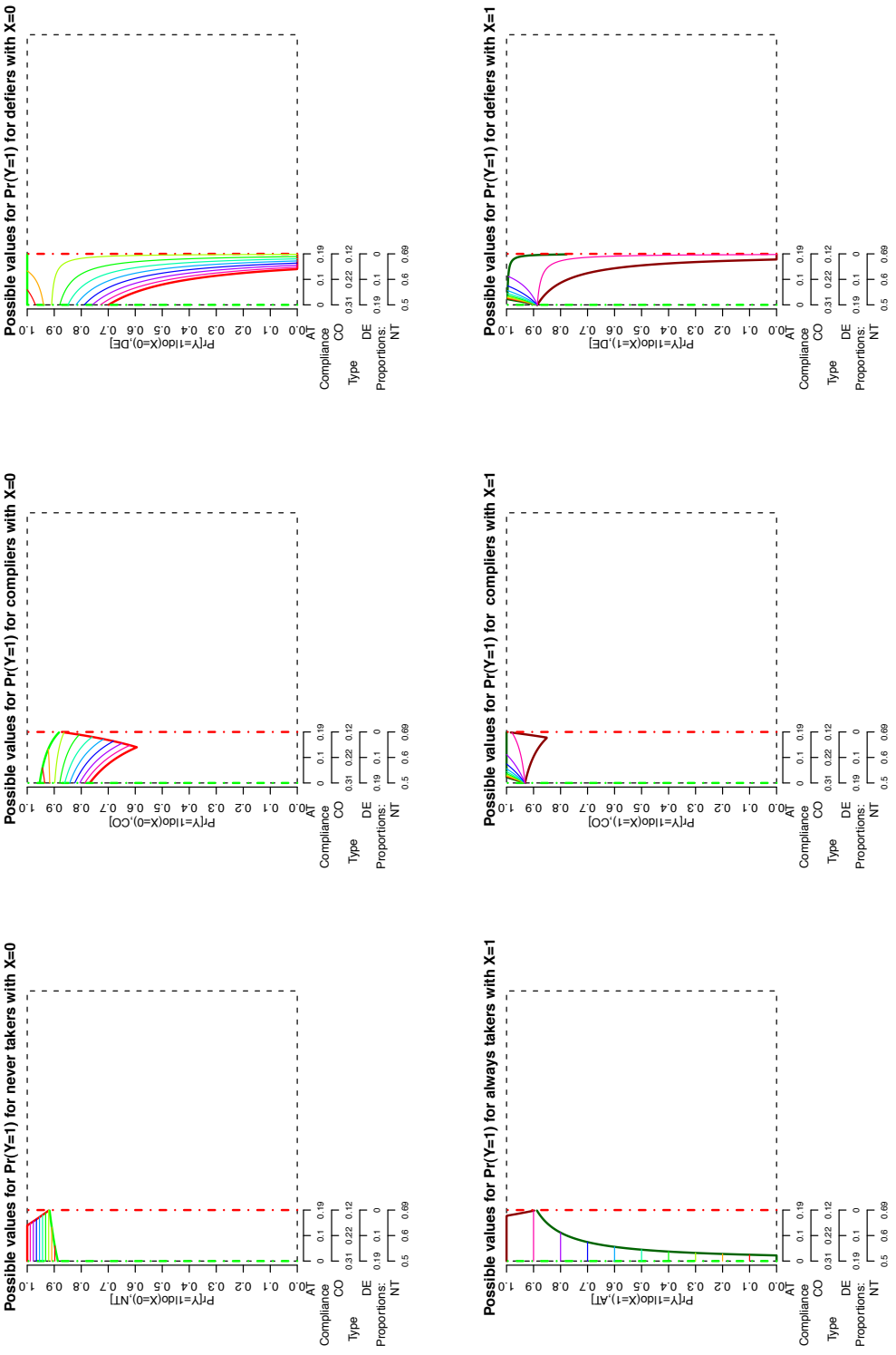

Figure 11: Depiction of the set of values for π_X vs. $\langle \gamma^0_{\mathrm{CO}}, \gamma^0_{\mathrm{DE}}, \gamma^0_{\mathrm{NT}} \rangle$ (upper row), and π_X vs. $\langle \gamma^1_{\mathrm{CO}}, \gamma^1_{\mathrm{DE}}, \gamma^1_{\mathrm{AT}} \rangle$ for the flu data.

Group	ACE Lower Bound	ACE Upper Bound
NT	$0 - u\gamma^0_{\mathrm{NT}}(\pi_X)$	$1 - l\gamma^0_{\mathrm{NT}}(\pi_X)$
CO	$l\gamma^1_{\mathrm{CO}}(\pi_X) - u\gamma^0_{\mathrm{CO}}(\pi_X)$ $= \gamma^1_{\mathrm{CO}}(\pi_X, u\gamma^1_{\mathrm{AT}}(\pi_X))$ $\quad -\gamma^0_{\mathrm{CO}}(\pi_X, l\gamma^0_{\mathrm{NT}}(\pi_X))$	$u\gamma^1_{\mathrm{CO}}(\pi_X) - l\gamma^0_{\mathrm{CO}}(\pi_X)$ $= \gamma^1_{\mathrm{CO}}(\pi_X, l\gamma^1_{\mathrm{AT}}(\pi_X))$ $\quad -\gamma^0_{\mathrm{CO}}(\pi_X, u\gamma^0_{\mathrm{NT}}(\pi_X))$
DE	$l\gamma^1_{\mathrm{DE}}(\pi_X) - u\gamma^0_{\mathrm{DE}}(\pi_X)$ $= \gamma^1_{\mathrm{DE}}(\pi_X, u\gamma^1_{\mathrm{AT}}(\pi_X))$ $\quad -\gamma^0_{\mathrm{DE}}(\pi_X, l\gamma^0_{\mathrm{NT}}(\pi_X))$	$u\gamma^1_{\mathrm{DE}}(\pi_X) - l\gamma^0_{\mathrm{DE}}(\pi_X)$ $= \gamma^1_{\mathrm{DE}}(\pi_X, l\gamma^1_{\mathrm{AT}}(\pi_X))$ $\quad -\gamma^0_{\mathrm{DE}}(\pi_X, u\gamma^0_{\mathrm{NT}}(\pi_X))$
AT	$l\gamma^1_{\mathrm{AT}}(\pi_X) - 1$	$u\gamma^1_{\mathrm{AT}}(\pi_X) - 0$
global	$p_{y_1,x_1\mid z_1} - p_{y_1,x_0\mid z_0}$ $\quad + \pi_{\mathrm{DE}} \cdot l\mathrm{ACE}_{\mathrm{DE}}(\pi_X) - \pi_{\mathrm{AT}}$ $= p_{y_1,x_1\mid z_0} - p_{y_1,x_0\mid z_1}$ $\quad + \pi_{\mathrm{CO}} \cdot l\mathrm{ACE}_{\mathrm{CO}}(\pi_X) - \pi_{\mathrm{AT}}$	$p_{y_1,x_1\mid z_1} - p_{y_1,x_0\mid z_0}$ $\quad + \pi_{\mathrm{DE}} \cdot u\mathrm{ACE}_{\mathrm{DE}}(\pi_X) + \pi_{\mathrm{NT}}$ $= p_{y_1,x_1\mid z_0} - p_{y_1,x_0\mid z_1}$ $\quad + \pi_{\mathrm{CO}} \cdot u\mathrm{ACE}_{\mathrm{CO}}(\pi_X) + \pi_{\mathrm{NT}}$

Table 4: Upper and Lower bounds on average causal effects for different groups, as a function of a feasible π_X. Here $\pi^c_{\mathrm{NT}} \equiv 1 - \pi_{\mathrm{NT}}$

(maximize) $\mathrm{ACE}_{\mathrm{NT}}$ and $\mathrm{ACE}_{\mathrm{AT}}$; this is an immediate consequence of the negative coefficients for γ^0_{NT} and γ^1_{AT} in the bounds for γ^x_{CO} and γ^x_{DE} in Table 2.

ACE bounds for the four compliance types are shown for the flu data in Figure 12. The ACE bounds for Compliers indicate that, under the observed distribution, the possibility of a zero ACE for Compliers is consistent with all feasible distributions over compliance types, except those for which the proportion of Defiers in the population is small.

Following [Pearl 2000; Robins 1989; Manski 1990; Robins and Rotnitzky 2004] we also consider the average causal effect on the entire population:

$$\mathrm{ACE}_{\mathrm{global}}(\pi_X, \{\gamma^x_{\mathbf{t}_X}\}) \equiv \sum_{\mathbf{t}_X \in \mathbb{D}_X} (\gamma^1_{\mathbf{t}_X}(\pi_X) - \gamma^0_{\mathbf{t}_X}(\pi_X))\pi_{\mathbf{t}_X};$$

upper and lower bounds taken over $\{\gamma^x_{\mathbf{t}_X}\}$ are defined similarly. The bounds given for $\mathrm{ACE}_{\mathbf{t}_X}$ in Table 5 are an immediate consequence of equations (8)–(11) which relate $p(y \mid x, z)$ to π_X and $\{\gamma^x_{\mathbf{t}_X}\}$. Before deriving the ACE bounds we need the following observation:

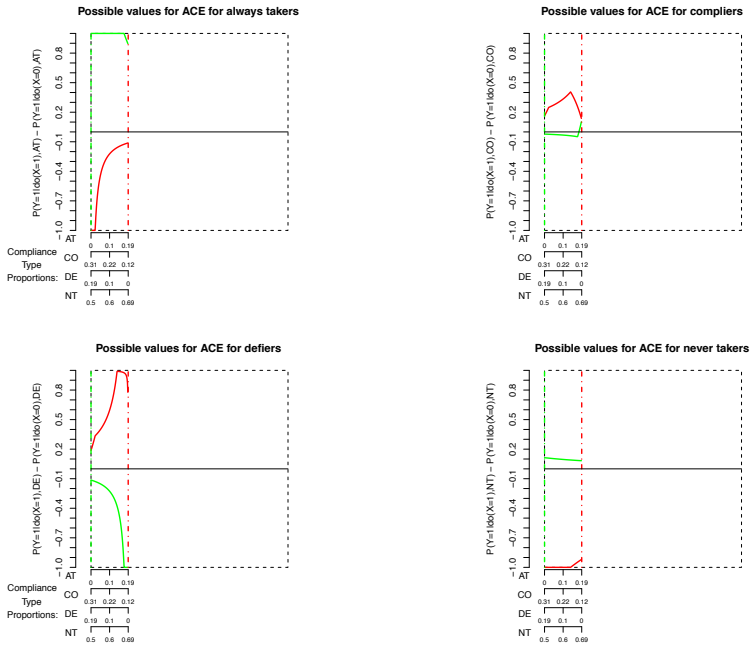

Figure 12: Depiction of the set of values for π_X vs. $\text{ACE}_{\mathbf{t}_X}(\pi_X)$ for $\mathbf{t}_X \in \mathbb{D}_X$ for the flu data.

Lemma 5 *For a given feasible* π_X *and* $p(y, x \mid z)$,

$$\begin{aligned}
&\text{ACE}_{\text{global}}(\pi_X, \{\gamma^x_{\mathbf{t}_X}\}) \\
&= p_{y_1,x_1|z_1} - p_{y_1,x_0|z_0} + \pi_{\text{DE}}(\gamma^1_{\text{DE}} - \gamma^0_{\text{DE}}) + \pi_{\text{NT}}\gamma^1_{\text{NT}} - \pi_{\text{AT}}\gamma^0_{\text{AT}} \qquad (21) \\
&= p_{y_1,x_1|z_0} - p_{y_1,x_0|z_1} + \pi_{\text{CO}}(\gamma^1_{\text{CO}} - \gamma^0_{\text{CO}}) + \pi_{\text{NT}}\gamma^1_{\text{NT}} - \pi_{\text{AT}}\gamma^0_{\text{AT}}. \qquad (22)
\end{aligned}$$

Proof: (21) follows from the definition of $\text{ACE}_{\text{global}}$ and the observation that $p_{y_1,x_1|z_1} = \pi_{\text{CO}}\gamma^1_{\text{CO}} + \pi_{\text{AT}}\gamma^1_{\text{AT}}$ and $p_{y_1,x_0|z_0} = \pi_{\text{CO}}\gamma^0_{\text{CO}} + \pi_{\text{NT}}\gamma^0_{\text{NT}}$. The proof of (22) is similar. □

Proposition 6 *For a given feasible* π_X *and* $p(y, x \mid z)$, *the compatible distribution which minimizes* [*maximizes*] $\text{ACE}_{\text{global}}$ *has*

$$\begin{aligned}
\langle \gamma^0_{\text{NT}}, \gamma^1_{\text{AT}} \rangle &= \langle l\gamma^0_{\text{NT}}, u\gamma^1_{\text{AT}} \rangle \quad [\langle u\gamma^0_{\text{NT}}, l\gamma^1_{\text{AT}} \rangle] \\
\langle \gamma^1_{\text{NT}}, \gamma^0_{\text{AT}} \rangle &= \langle 0, 1 \rangle \qquad\qquad [\langle 1, 0 \rangle]
\end{aligned}$$

thus also minimizes [*maximizes*] ACE_{CO} *and* ACE_{DE}, *and conversely maximizes* [*minimizes*] ACE_{AT} *and* ACE_{NT}.

Proof: The claims follow from equations (21) and (22), together with the fact that γ^0_{AT} and γ^1_{NT} are unconstrained, so $\text{ACE}_{\text{global}}$ is minimized by taking $\gamma^0_{\text{AT}} = 1$ and $\gamma^1_{\text{NT}} = 0$, and maximized by taking $\gamma^0_{\text{AT}} = 0$ and $\gamma^1_{\text{NT}} = 1$. □

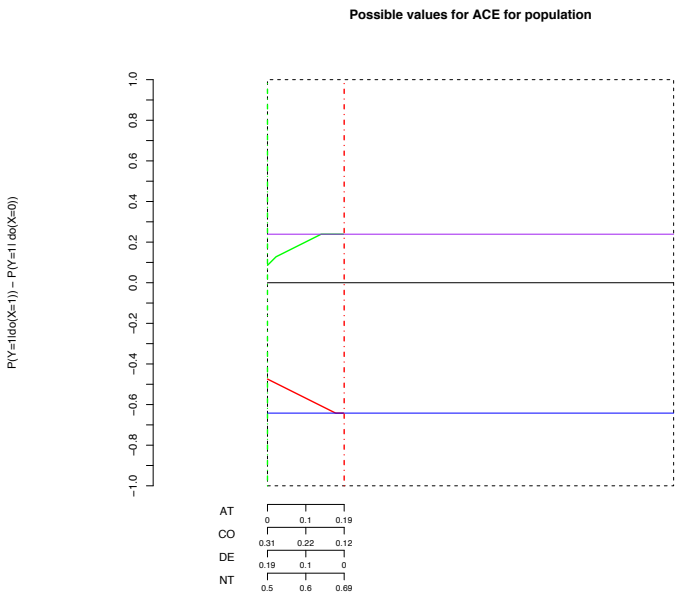

Figure 13: Depiction of the set of values for π_X vs. the global ACE for the flu data. The horizontal lines represent the overall bounds on the global ACE due to Pearl.

It is of interest here that although the definition of $\text{ACE}_{\text{global}}$ treats the four compliance types symmetrically, the compatible distribution which minimizes [maximizes] this quantity (for a given π_X) does not: it always corresponds to the scenario in which the treatment has the smallest [greatest] effect on Compliers and Defiers.

The bounds on the global ACE for the flu vaccine data of [Hirano, Imbens, Rubin, and Zhou 2000] are shown are shown in Figure 13.

Finally we note that it would be simple to develop similar bounds for other measures such as the Causal Relative Risk and Causal Odds Ratio.

6 Instrumental inequalities

The expressions involved in the upper bound on π_{AT} in (16) appear similar to those which occur in Pearl's instrumental inequalities. Here we show that the requirement that $\mathcal{P}_X \neq \emptyset$, or equivalently, $l\pi_{\text{AT}} \leq u\pi_{\text{AT}}$ is in fact equivalent to the instrumental inequality. This also provides an interpretation as to what may be inferred from the violation of a specific inequality.

Theorem 7 *The following conditions place equivalent restrictions on* $p(x \mid z)$ *and* $p(y \mid x = 0, z)$*:*

(a1) $\max\{0,\ p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} \leq$

$$\min\Big\{1-\textstyle\sum_j p(y{=}j, x{=}0 \mid z{=}j),\ 1-\sum_k p(y{=}k, x{=}0 \mid z{=}1{-}k)\Big\};$$

(a2) $\max\Big\{\sum_j p(y{=}j, x{=}0 \mid z{=}j),\ \sum_k p(y{=}k, x{=}0 \mid z{=}1-k)\Big\} \leq 1.$

Similarly, the following place equivalent restrictions on $p(x \mid z)$ *and* $p(y \mid x{=}1, z)$*:*

(b1) $\max\{0, p(x=1 \mid z=0) + p(x=1 \mid z=1) - 1\} \leq$
$$\min\Big\{\textstyle\sum_j p(y{=}j, x{=}1 \mid z{=}j),\ \sum_k p(y{=}k, x{=}1 \mid z{=}1{-}k)\Big\};$$

(b2) $\max\Big\{\sum_j p(y{=}j, x{=}1 \mid z{=}j),\ \sum_k p(y{=}k, x{=}1 \mid z{=}1-k)\Big\} \leq 1.$

Thus the instrumental inequality (a2) corresponds to the requirement that the upper bounds on $p(\text{AT})$ resulting from $p(x \mid z)$ and $p(y=1 \mid x=0, z)$ be greater than the lower bound on $p(\text{AT})$ (derived solely from $p(x \mid z)$). Similarly for (b2) and the upper bounds on $p(\text{AT})$ resulting from $p(y{=}1 \mid x{=}1, z)$.

Proof: [(a1) $\Leftrightarrow$ (a2)] We first note that:

$$\begin{aligned}
&1-\textstyle\sum_j p(y{=}j, x{=}0 \mid z{=}j) \geq \Big(\sum_j p(x{=}1 \mid z{=}j)\Big) - 1 \\
&\quad\Leftrightarrow\ \textstyle\sum_j (1 - p(y{=}j, x{=}0 \mid z{=}j)) \geq \sum_j p(x{=}1 \mid z{=}j) \\
&\quad\Leftrightarrow\ \textstyle\sum_j (p(y{=}1-j, x{=}0 \mid z{=}j) + p(x{=}1 \mid z{=}j)) \geq \sum_j p(x{=}1 \mid z{=}j) \\
&\quad\Leftrightarrow\ \textstyle\sum_j p(y{=}j, x{=}0 \mid z{=}j) \geq 0.
\end{aligned}$$

which always holds. By a symmetric argument we can show that it always holds that:

$$1-\textstyle\sum_j p(y{=}j, x{=}0 \mid z{=}1-j) \geq \Big(\sum_j p(x{=}1 \mid z{=}j)\Big) - 1.$$

Thus if (a1) does not hold then $\max\{0,\ p(x{=}1 \mid z{=}0) + p(x{=}1 \mid z{=}1) - 1\} = 0$. It is then simple to see that (a1) does not hold iff (a2) does not hold.

[(b1) $\Leftrightarrow$ (b2)] It is clear that neither of the sums on the RHS of (b1) are negative, hence if (b1) does not hold then $\max\{0, p(x = 1 \mid z = 0) + p(x = 1 \mid z = 1) - 1\} = \Big(\sum_j p(x{=}1 \mid z{=}j)\Big) - 1$. Now

$$\begin{aligned}
&\textstyle\sum_j p(y{=}j, x{=}1 \mid z{=}j) < \Big(\sum_j p(x{=}1 \mid z{=}j)\Big) - 1 \\
&\quad\Leftrightarrow\ 1 < \textstyle\sum_j p(y{=}j, x{=}1 \mid z{=}1-j).
\end{aligned}$$

Likewise

$$\begin{aligned}
&\textstyle\sum_j p(y{=}j, x{=}1 \mid z{=}1-j) < \Big(\sum_j p(x{=}1 \mid z{=}j)\Big) - 1 \\
&\quad\Leftrightarrow\ 1 < \textstyle\sum_j p(y{=}j, x{=}1 \mid z{=}j).
\end{aligned}$$

Thus (b1) fails if and only if (b2) fails. □

This equivalence should not be seen as surprising since [Bonet 2001] states that the instrument inequalities (a2) and (b2) are sufficient for a distribution to be compatible with the binary IV model. This is not the case if, for example, X takes more than 2 states.

6.1 Which alternatives does a test of the instrument inequalities have power against?

[Pearl 2000] proposed testing the instrument inequalities (a2) and (b2) as a means of testing the IV model; [Ramsahai 2008] develops tests and analyzes their properties. It is then natural to ask what should be inferred from the failure of a specific instrumental inequality. It is, of course, always possible that randomization has failed. If randomization is not in doubt, then the exclusion restriction (1) must have failed in some way. The next result implies that tests of the inequalities (a2) and (b2) have power, respectively, against failures of the exclusion restriction for Never Takers (with $X = 0$) and Always Takers (with $X = 1$):

Theorem 8 *The conditions* (RX), ($\mathrm{RY}_{X=0}$) *and* ($\mathrm{E}_{X=0}$) *described below imply* (a2)*; similarly* (RX), ($\mathrm{RY}_{X=1}$) *and* ($\mathrm{E}_{X=1}$) *imply* (b2)*.*

$$
\begin{aligned}
&(\mathrm{RX}) && Z \perp\!\!\!\perp \mathfrak{t}_X \text{ equivalently } Z \perp\!\!\!\perp X_{z=0}, X_{z=1} : \\
&(\mathrm{RY}_{X=0}) && Z \perp\!\!\!\perp Y_{x=0,z=0} \mid \mathfrak{t}_X = \mathrm{NT}; \quad Z \perp\!\!\!\perp Y_{x=0,z=1} \mid \mathfrak{t}_X = \mathrm{NT}; \\
&(\mathrm{RY}_{X=1}) && Z \perp\!\!\!\perp Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT}; \quad Z \perp\!\!\!\perp Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT}; \\
&(\mathrm{E}_{X=0}) && p(Y_{x=0,z=0} = Y_{x=0,z=1} \mid \mathfrak{t}_X = \mathrm{NT}) = 1; \\
&(\mathrm{E}_{X=1}) && p(Y_{x=1,z=0} = Y_{x=1,z=1} \mid \mathfrak{t}_X = \mathrm{AT}) = 1.
\end{aligned}
$$

Conditions (RX) and ($\mathrm{RY}_{X=x}$) correspond to the assumption of randomization with respect to compliance type and response type. For the purposes of technical clarity we have stated condition ($\mathrm{RY}_{X=x}$) in the weakest form possible. However, we know of no subject matter knowledge which would lead one to believe that (RX) and ($\mathrm{RY}_{X=x}$) held, without also implying the stronger assumption (2). In contrast, the exclusion restrictions ($\mathrm{E}_{X=x}$) are significantly weaker than (1), e.g. one could conceive of situations where assignment had an effect on the outcome for Always Takers, but not for Compliers. It should be noted that tests of the instrument inequalities have no power to detect failures of the exclusion restriction for Compliers or defier.

We first prove the following Lemma, which also provides another characterization of the instrument inequalities:

Lemma 9 *Suppose* (RX) *holds and* $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \mathrm{NT}$ *then* (a2) *holds. Similarly, if* (RX) *holds and* $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \mathrm{AT}$ *then* (b2) *holds.*

Note that the conditions in the antecedent make no assumption regarding the existence of counterfactuals for Y.

Proof: We prove the result for Never Takers; the other proof is similar. By hypothesis we have:

$$p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \mathrm{NT}) = p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \mathrm{NT}) \;\equiv\; \gamma^0_{\mathrm{NT}}. \tag{23}$$

In addition,

$$\begin{aligned}
&p(Y = 1 \mid Z = 0, X = 0) \\
&\quad = \; p(Y = 1 \mid Z = 0, X = 0, X_{z=0} = 0) \\
&\quad = \; p(Y = 1 \mid Z = 0, X_{z=0} = 0) \\
&\quad = \; p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \mathrm{CO})\, p(\mathfrak{t}_X = \mathrm{CO} \mid Z = 0, X_{z=0} = 0) \\
&\qquad\quad + p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \mathrm{NT})\, p(\mathfrak{t}_X = \mathrm{NT} \mid Z = 0, X_{z=0} = 0) \\
&\quad = \; p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \mathrm{CO})\, p(\mathfrak{t}_X = \mathrm{CO} \mid X_{z=0} = 0) \\
&\qquad\quad + \gamma^0_{\mathrm{NT}}\, p(\mathfrak{t}_X = \mathrm{NT} \mid X_{z=0} = 0).
\end{aligned} \tag{24}$$

The first three equalities here follow from consistency, the definition of the compliance types and the law of total probability. The final equality uses (RX). Similarly, it may be shown that

$$\begin{aligned}
&p(Y = 1 \mid Z = 1, X = 0) \\
&\quad = \; p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \mathrm{DE}) p(\mathfrak{t}_X = \mathrm{DE} \mid X_{z=1} = 0) \\
&\qquad\quad + \gamma^0_{\mathrm{NT}}\, p(\mathfrak{t}_X = \mathrm{NT} \mid X_{z=1} = 0).
\end{aligned} \tag{25}$$

Equations (24) and (25) specify two averages of three quantities, thus taking $u = p(Y = 1 \mid Z = 0, \mathfrak{t}_X = \mathrm{CO})$, $v = p(Y = 1 \mid Z = 1, \mathfrak{t}_X = \mathrm{DE})$ and $w = \gamma^0_{\mathrm{NT}}$, we may apply the analysis of §2.3. This then leads to the upper bound on π_{AT} given by equation (15). (Note that the lower bounds on π_{AT} are derived from $p(x \mid z)$ and hence are unaffected by dropping the exclusion restriction.) The requirement that there exist some feasible distribution π_X then implies equation (a2) which is shown in Theorem 7 to be equivalent to (b2) as required. □

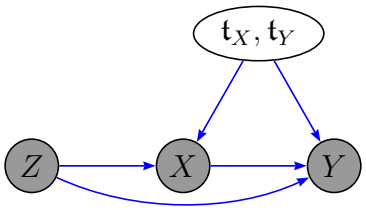

Figure 14: Graphical representation of the model given by the randomization assumption (2) alone. It is no longer assumed that Z does not have a direct effect on Y.

Proof of Theorem 8: We establish that (RX), (RY$_{X=0}$), (E$_{X=0}$) $\Rightarrow$ (a2). The proof of the other implication is similar. By Lemma 9 it is sufficient to establish that $Y \perp\!\!\!\perp Z \mid \mathfrak{t}_X = \text{NT}$.

$$
\begin{aligned}
& p(Y=1 \mid Z=0, \mathfrak{t}_X = \text{NT}) \\
&= p(Y=1 \mid Z=0, X=0, \mathfrak{t}_X = \text{NT}) && \text{definition of NT;} \\
&= p(Y_{x=0,z=0}=1 \mid Z=0, X=0, \mathfrak{t}_X = \text{NT}) && \text{consistency;} \\
&= p(Y_{x=0,z=0}=1 \mid Z=0, \mathfrak{t}_X = \text{NT}) && \text{definition of NT;} \\
&= p(Y_{x=0,z=0}=1 \mid \mathfrak{t}_X = \text{NT}) && \text{by (RY}_{X=0}\text{);} \\
&= p(Y_{x=0,z=1}=1 \mid \mathfrak{t}_X = \text{NT}) && \text{by (E}_{X=0}\text{);} \\
&= p(Y_{x=0,z=1}=1 \mid Z=1, \mathfrak{t}_X = \text{NT}) && \text{by (RY}_{X=0}\text{);} \\
&= p(Y=1 \mid Z=1, \mathfrak{t}_X = \text{NT}) && \text{consistency, NT.}
\end{aligned}
$$

□

A similar result is given in [Cai, Kuroki, Pearl, and Tian 2008], who consider the *Average Controlled Direct Effect*, given by:

$$\text{ACDE}(x) \equiv p(Y_{x,z=1}=1) \; - \; p(Y_{x,z=0}=1),$$

under the model given solely by the equation (2), which corresponds to the graph in Figure 14. Cai *et al.* prove that under this model the following bounds obtain:

$$\text{ACDE}(x) \;\geq\; p(y=0, x \mid z=0) + p(y=1, x \mid z=1) - 1, \tag{26}$$

$$\text{ACDE}(x) \;\leq\; 1 - p(y=0, x \mid z=1) - p(y=1, x \mid z=0). \tag{27}$$

It is simple to see that ACDE(x) will be bounded away from 0 for some x iff one of the instrumental inequalities is violated. This is as we would expect: the IV model of Figure 1 is a sub-model of Figure 14, but if ACDE(x) is bounded away from 0 then the $Z \rightarrow Y$ edge is present, and hence the exclusion restriction (1) is incompatible with the observed distribution.

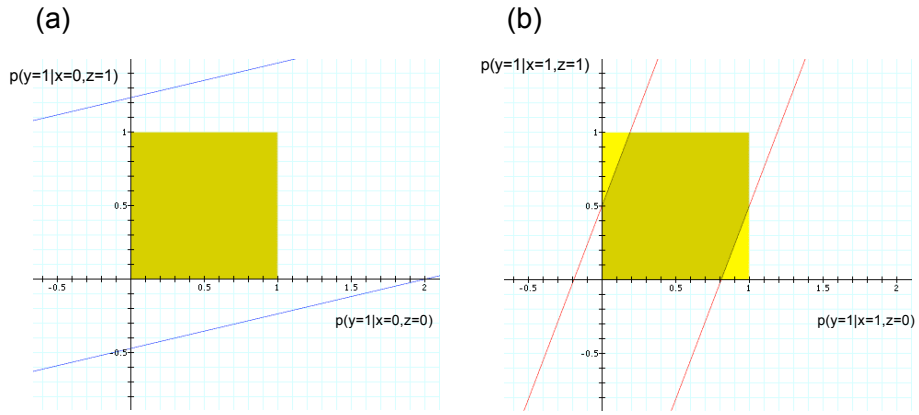

Figure 15: Illustration of the possible values for $p(y \mid x, z)$ compatible with the instrument inequalities, for a given distribution $p(x|z)$.The darker shaded region satisfies the inequalities: (a) $X = 0$, inequalities (a2); (b) $X = 1$, inequalities (b2). In this example $p(x{=}1 \mid z{=}0) = 0.84$, $p(x{=}1 \mid z{=}1) = 0.32$. Since $0.84/(1 - 0.32) > 1$, (a2) is trivially satisfied; see proof of Theorem 10.

6.2 How many instrument inequalities may be violated by a single distribution?

Theorem 10 *For any distribution $p(x, y \mid z)$, at most one of the four instrument inequalities:*

(a2.1) $\sum_j p(y{=}j, x{=}0 \mid z{=}j) \leq 1$; (a2.2) $\sum_j p(y{=}j, x{=}0 \mid z{=}1{-}j) \leq 1$;

(b2.1) $\sum_j p(y{=}j, x{=}1 \mid z{=}j) \leq 1$; (b2.2) $\sum_j p(y{=}j, x{=}1 \mid z{=}1{-}j) \leq 1$;

is violated.

Proof: We first show that at most one of (a2.1) and (a2.2) may be violated. Letting $\theta_{ij} = p(y = 1 \mid x = j, z = i)$ we may express these inequalities as:

$$\begin{aligned} \theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} &\leq p_{x_1|z_0}, \qquad \text{(a2.1)} \\ \theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} &\geq -p_{x_1|z_1}, \qquad \text{(a2.2)} \end{aligned}$$

giving two half-planes in $(\theta_{00}, \theta_{10})$-space (see Figure 15(a)). Since the lines defining the half-planes are parallel, it is sufficient to show that the half-planes always intersect, and hence that the regions in which (a2.1) and (a2.2) are violated are disjoint. However, this is immediate since the (non-empty) set of points for which $\theta_{10} \cdot p_{x_0|z_1} - \theta_{00} \cdot p_{x_0|z_0} = 0$ always satisfy both inequalities.

The proof that at most one of (b2.1) and (b2.2) may be violated is symmetric.

We now show that the inequalities (a2.1) and (a2.2) place non-trivial restrictions on $(\theta_{00}, \theta_{10})$ iff (b2.1) and (b2.2) place trivial restrictions on $(\theta_{01}, \theta_{11})$. The line corresponding

to (a2.1) passes through $(\theta_{00}, \theta_{10}) = (-p_{x_1|z_0}/p_{x_0|z_0}, 0)$ and $(0, p_{x_1|z_0}/p_{x_0|z_1})$; since the slope of the line is non-negative, it has non-empty intersection with $[0,1]^2$ iff $p_{x_1|z_0}/p_{x_0|z_1} \leq 1$. Thus there are values of $(\theta_{01}, \theta_{11}) \in [0,1]^2$ which fail to satisfy (a2.1) iff $p_{x_1|z_0}/p_{x_0|z_1} < 1$. By a similar argument it may be shown that (a2.2) is non-trivial iff $p_{x_1|z_1}/p_{x_0|z_0} < 1$, which is equivalent to $p_{x_1|z_0}/p_{x_0|z_1} < 1$.

The proof is completed by showing that (b2.1) and (b2.2) are non-trivial if and only if $p_{x_1|z_0}/p_{x_0|z_1} > 1$. □

Corollary 11 *Every distribution $p(x, y \mid z)$ is consistent with randomization* (RX) *and* (2)*, and at least one of the exclusion restrictions* $\mathrm{E}_{X=0}$ *or* $\mathrm{E}_{X=1}$.

6.2.1 Flu Data Revisited

For the data in Table 3, all of the instrument inequalities hold. Consequently there is no evidence of a direct effect of Z on Y. (Again we emphasize that unlike [Hirano, Imbens, Rubin, and Zhou 2000], we are not using any information on baseline covariates in the analysis.) Finally we note that, since all of the instrumental inequalities hold, maximum likelihood estimates for the distribution $p(x, y \mid z)$ under the IV model are given by the empirical distribution. However, if one of the IV inequalities were to be violated then the MLE would not be equal to the empirical distribution, since the latter would not be a law within the IV model. In such a circumstance a fitting procedure would be required; see [Ramsahai 2008, Ch. 5].

7 Conclusion

We have built upon and extended the work of Pearl, displaying how the range of possible distributions over types compatible with a given observed distribution may be characterized and displayed geometrically. Pearl's bounds on the global ACE are sometimes objected to on the grounds that they are too extreme, since for example, the upper bound presupposes a 100% success rate among Never Takers if they were somehow to receive treatment, likewise a 100% failure rate among Always Takers were they not to receive treatment. Our analysis provides a framework for performing a sensitivity analysis. Lastly, our analysis relates the IV inequalities to the bounds on direct effects.

Acknowledgements

This research was supported by the U.S. National Science Foundation (CRI 0855230) and U.S. National Institutes of Health (R01 AI032475) and Jesus College, Oxford where Thomas

Richardson was a Visiting Senior Research Fellow in 2008. The authors used Avitzur's *Graphing Calculator* software (`www.pacifict.com`) to construct two and three dimensional plots. We thank McDonald, Hiu and Tierney for giving us permission to use their flu vaccine data.

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 17th 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.

Erosheva, E. A. (2005). Comparing latent structures of the Grade of Membership, Rasch, and latent class models. *Psychometrika 70*, 619–628.

Fienberg, S. E. and J. P. Gilbert (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association 65*, 694–701.

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.

Manski, C. (1990). Non-parametric bounds on treatment effects. *American Economic Review 80*, 351–374.

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.

Robins, J. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A focus on AIDS*. Washington, D.C.: U.S. Public Health Service.

Robins, J. and A. Rotnitzky (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika 91*(4), 763–783.