BAYESIAN STATISTICS 9, pp. 569–610.
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, 2011

Transparent Parametrizations of Models for Potential Outcomes

Thomas S. Richardson, Robin J. Evans
University of Washington, USA
`thomasr@uw.edu`, `rje42@uw.edu`

James M. Robins
Harvard School of Public Health, USA
`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 an example that the resulting posteriors may be highly sensitive to the specification of the prior distribution over compliance types. To address this, we present several novel "transparent" re-parametrizations of the likelihood that separate the identified and non-identified parts of the parameter. In addition, we develop parametrizations that are robust to model mis-specification under the "intent-to-treat" null hypothesis that Z and Y are independent.

Keywords and Phrases: Bounds; Continuous Covariates; Exclusion Restriction; Instrumental Inequality; ITT-Null-Robust; Model Mis-Specification; Parametrization; Prior Sensitivity.

1. INTRODUCTION

The potential outcomes model for causal inference is a well-established framework for formalizing causal assumptions and modelling causal effects; see Neyman (1923); Rubin (1974). However, in many contexts, the causal estimands of interest are not identified by the observed data. Even in the asymptotic limit, there may remain a range of values for a parameter of interest that are logically possible, rather than a single point. Such a parameter is *partially identified*, and is *entirely non-identified* if, in the limit, the data impose no restriction at all.

It is often argued that identifiability is of secondary importance in a Bayesian analysis provided that the prior and likelihood lead to a proper joint posterior for all the parameters in the model. Following Leamer (1978), Gustafson (2005) and Greenland (2005), we argue that partially identified models should be re-parameterized so that the complete parameter vector may be divided into point-identified and entirely non-identified subvectors. 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, in which the "instrument" Z, the exposure measure X, and the final response Y are all binary. We then extend this approach to the analysis of a randomized encouragement design, though still with binary treatment and response, under a variety of different assumptions.

In Section 5, we develop novel variation independent smooth parametrizations that permit this approach to be applied in the context of continuous or discrete baseline covariates. As the response Y is binary, these parametrizations are of necessity complex and somewhat non-intuitive.

In Section 6, we consider a successfully randomized controlled trial in which the following both hold: (i) the random assignment of Z and (ii) the exclusion restriction that Z has no effect on Y except through its effect on X. The exclusion restriction is guaranteed to hold in a placebo-controlled double-blind trial in which the active treatment is which the active treatment is without side-effects and unavailable to patients in the control arm. Under this model, the sharp null hypothesis of no causal effect of X on Y implies the Intent-To-Treat (ITT) null hypothesis that Z and Y are independent both marginally and conditional on the baseline covariates V. We provide a second transparent parametrization with the following important robustness property: under the ITT null, a Bayes estimator of the conditional covariance of Z and Y given V converges to zero, even when both the estimates of the distribution $p(x \mid y, z, v)$ and of $p(y \mid z = 0, v)$ are inconsistent due to model mis-specification. This is important because mis-specification is inevitable whenever V has continuous components.

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 the induced model for the joint distribution of the observed data implied by each model. In Section 5, we extend the approach to incorporate baseline covariates. In Section 6, we modify the approach presented in Section 5 to preserve consistency under the ITT null.

2. BASIC CONCEPTS

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

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 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; see Imbens and Rubin (1997).*

$X_{z=0}$	$X_{z=1}$	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*, 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).

2.1. *Notation*

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

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

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

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

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

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

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

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.

2.2. *Randomization Assumption*

We will make the randomization assumption that the distribution of types $(\mathsf{t}_X, \mathsf{t}_Y)$ 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.

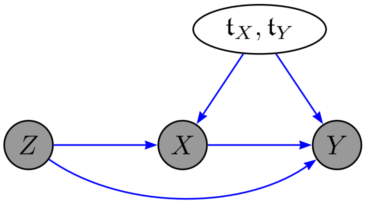


Figure 1: *Graphical representation of the model given by assumption* (1). *The shaded nodes are observed. In this model* t_X *takes* 4 *states, while* t_Y *takes* 16.

3. A SIMPLE MOTIVATING EXAMPLE

Pearl (2000) and Chickering and Pearl (2000) use potential outcomes to analyze the data in Table 2, which arise from a double-blind placebo-controlled randomized trial of Cholestyramine; see Efron 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.

Table 2: *Lipid/Cholestyramine data; originally considered by Efron and Freeman (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

Table 3: *Response types under the exclusion restriction (2); see Heckerman and Shachter (1995).*

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

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 double-blind 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 3. 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 $\gamma^{i\cdot}_{\mathfrak{t}_X} \equiv P(Y_{x=i\cdot} = 1 \mid \mathfrak{t}_X)$.

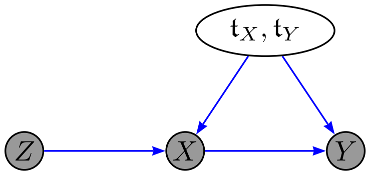


Figure 2: *Graphical representation of the IV model given by assumptions* (2) *and* (1)*. In this model* $\mathfrak{t}_X$ *takes* 4 *states, while* $\mathfrak{t}_Y$ *takes* 4*.*

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.

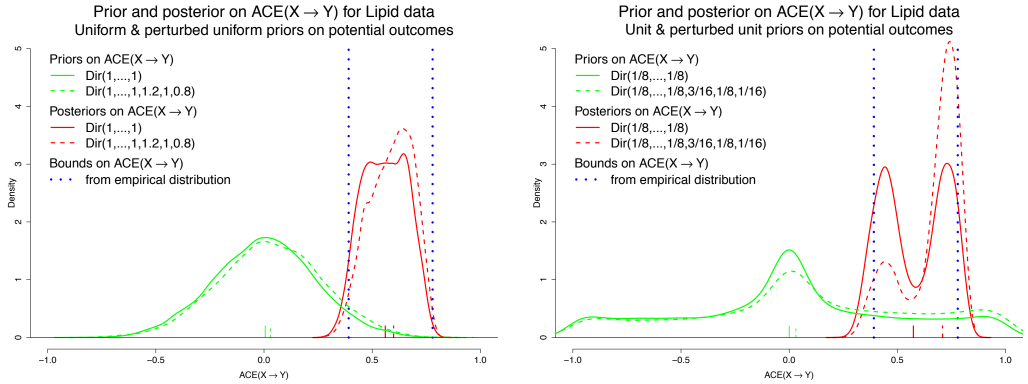


Figure 3: *Prior to posterior analysis for* $\text{ACE}(X \to Y)$ *for the Lipid data; priors are green; posteriors are red; vertical lines indicate bounds on the ACE evaluated at the empirical distribution. Tick marks indicate respective medians. See text for further details.*

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 $\text{Dir}(1, \dots, 1)$ 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 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$(X \to Y)$, 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 patients who are of type "Helped" (rather than "Never Recover") is directly relevant to ACE$(X \to Y)$.

3.1. *Separating the Identified from the Unidentified*

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

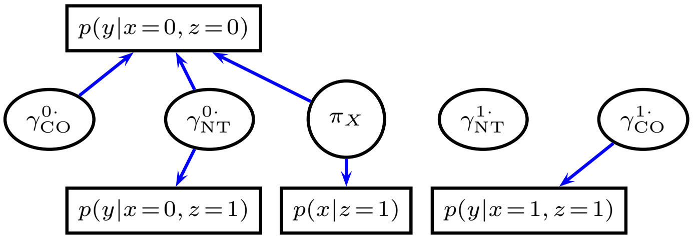

$p(y|x=0, z=0)$
$\gamma^{0\cdot}_{\text{CO}}$
$\gamma^{0\cdot}_{\text{NT}}$
π_X
$\gamma^{1\cdot}_{\text{NT}}$
$\gamma^{1\cdot}_{\text{CO}}$
$p(y|x=0, z=1)$
$p(x|z=1)$
$p(y|x=1, z=1)$

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.*

The parameters π_X, $\gamma^{1\cdot}_{\text{CO}}$, $\gamma^{0\cdot}_{\text{CO}}$, and $\gamma^{0\cdot}_{\text{NT}}$ are identified thus:

$$\pi_{\text{CO}} = p_{x_1|z_1}, \quad \gamma^{1\cdot}_{\text{CO}} = p_{y_1|x_1,z_1}, \quad \gamma^{0\cdot}_{\text{CO}} = (p_{y_1,x_0|z_0} - p_{y_1,x_0|z_1})/p_{x_1|z_1},$$

$$\pi_{\text{NT}} = p_{x_0|z_1}, \qquad \gamma^{0\cdot}_{\text{NT}} = p_{y_1|x_0,z_1}.$$

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

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

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, which is a consequence of the empirical distribution being close to violating (3). (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 although (2) implies (3), the converse does not hold. Similarly, if the posterior probability of (3) holding is high, this does not imply that the posterior probability of (2) is high, unless there is high prior conditional probability that (2) is true given that (3) is true. This follows from the fact that the posterior probability that (2) is true given that (3) is true is equal to the conditional prior probability that (2) is true given that (3) is true. The model that allows (2) to be violated is of the same dimension as model (2).

In this example, where Equation (2) is assumed to hold, we could have used $(\pi_X, \gamma^{1\cdot}_{\mathrm{CO}}, \gamma^{0\cdot}_{\mathrm{CO}}, \gamma^{0\cdot}_{\mathrm{NT}})$ 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 $\gamma^{i\cdot}_{\mathfrak{t}_X}$ may themselves be partially identified; see Richardson and Robins (2010).

3.2. *Posterior Distributions for the ACE*

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

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

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

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 $(\mathfrak{t}_X, \mathfrak{t}_Y)$.

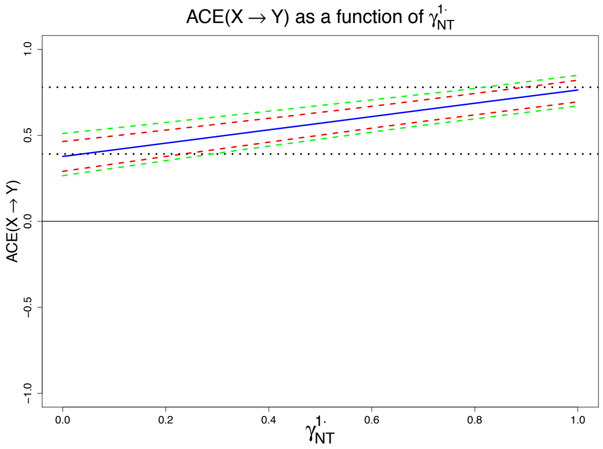


Figure 5: *The posterior for the* ACE$(X \to Y)$ *for the Lipid data displayed as a function of the (completely) unidentified parameter* $\gamma_{\mathrm{NT}}^{1\cdot}$*:* (blue) *posterior median;* (red) 2.5% *and* 97.5% *quantiles;* (green) *simultaneous* 95% *posterior region obtained from a* 95% *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)*.*

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*: $\gamma_{\mathrm{NT}}^{01} = \gamma_{\mathrm{NT}}^{00}$, so among Never Takers the distributions of Y_{00} and Y_{01} are the same.

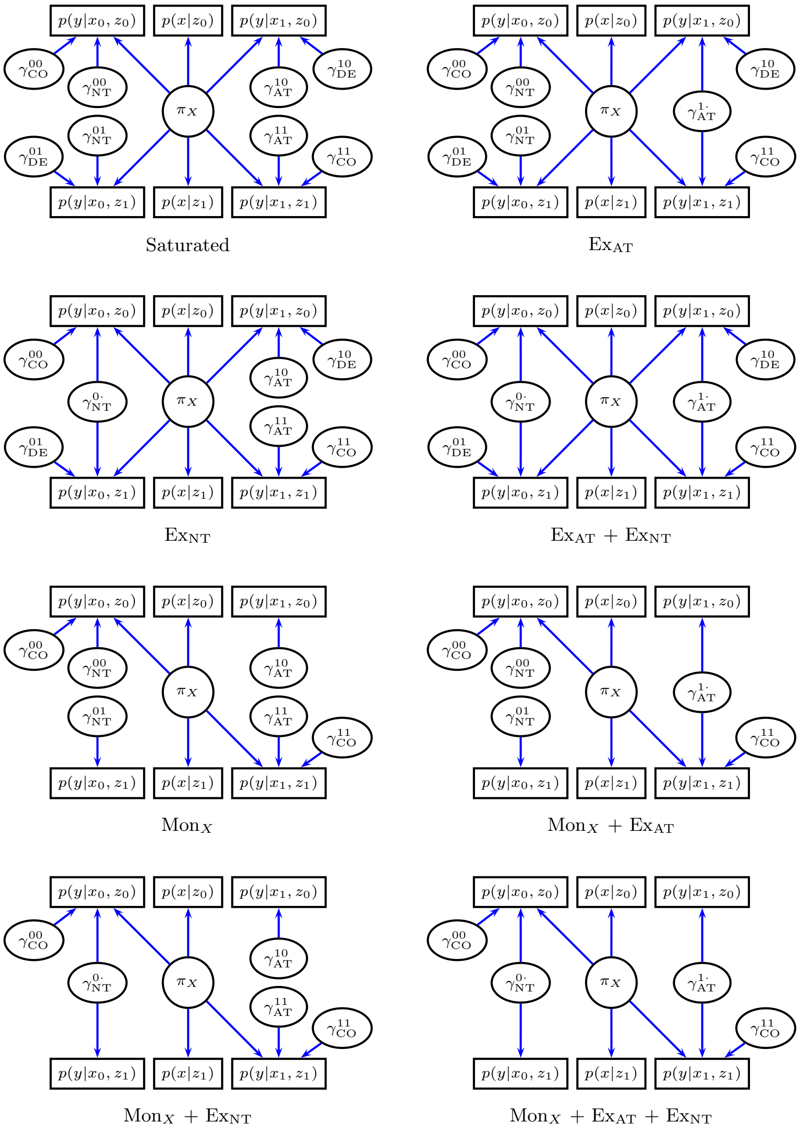
(Ex$_{\mathrm{AT}}$) *Stochastic exclusion for* AT *under exposure*: $\gamma_{\mathrm{AT}}^{11} = \gamma_{\mathrm{AT}}^{10}$, so among Always Takers the distributions of Y_{10} and Y_{11} are the same.

Note that assumption (2) implies stochastic exclusion for all compliance types under all exposures, *i.e.*, $\gamma_{\mathfrak{t}_X}^{ij} = \gamma_{\mathfrak{t}_X}^{ij'}$ for all $i, j, j' \in \{0,1\}$ and all $\mathfrak{t}_X \in \mathbb{D}_X$. Figure 6 and Table 4 list these eight models. Imposing other exclusion restrictions, besides Ex$_{\mathrm{AT}}$ or Ex$_{\mathrm{NT}}$, will correspond to merely relabeling a single node $\gamma_{\mathfrak{t}_X}^{ij}$ in Figure 6 with $\gamma_{\mathfrak{t}_X}^{i\cdot}$. 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.

4.1. *Inequalities Defining Models with Defiers*

Results of Balke and Pearl (1997) and Bonet (2001) imply that the set of distributions arising from a potential outcomes model satisfying (1), Ex$_{\mathrm{AT}}$ and Ex$_{\mathrm{NT}}$ may


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

Saturated

Ex_{AT}

Ex_{NT}

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

Mon_X

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

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

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

Figure 6: *Functional dependencies in the eight models. Terms $\gamma^{ij}_{\mathbf{t}_X}$ that do not appear in the likelihood are not shown. See also Table 4.*

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 relabeling 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):

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

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

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

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

where $\mathrm{ACDE}_{\mathfrak{t}_X}(x) \equiv E[Y_{x1} - Y_{x0} \mid \mathfrak{t}_X]$. Causal contrasts such as $\mathrm{ACDE}_{\mathrm{NT}}(x_0)$ and $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$ were introduced in Robins (1986, Section 12.2), in the context of estimating treatment in the presence of censoring by competing causes of death. Rubin (1998, 2004) and Frangakis and Rubin (2002) later coined the term "principal stratum direct effect". Bounds for $\mathrm{ACDE}_{\mathrm{AT}}$ and $\mathrm{ACDE}_{\mathrm{NT}}$ have been derived by Zhang and Rubin (2003); Hudgens *et al.* (2003) and Imai (2008).

$\mathrm{ACDE}(x_0)$ may be bounded away from 0 iff $\mathrm{ACDE}_{\mathrm{NT}}(x_0)$ may be bounded away from 0 in the same direction (hence $\mathrm{Ex}_{\mathrm{NT}}$ does not hold); see Kaufman *et al.* (2009) and Cai *et al.* (2008). Likewise, with $\mathrm{ACDE}(x_1)$, $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$ and $\mathrm{Ex}_{\mathrm{AT}}$. 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 $\mathrm{Ex}_{\mathrm{AT}}$ or $\mathrm{Ex}_{\mathrm{NT}}$ (or both).

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

4.2. *Weaker Assumptions*

It is worth noting that the inequalities (4) and (5) are implied by the larger counterfactual model which only assumes:

$$p(Y_{x=i,z=0} = 1) \;=\; p(Y_{x=i,z=1} = 1), \quad \text{for } i = 0, 1, \tag{7}$$

and

$$Z \perp\!\!\!\perp Y_{x=0,z} \quad \text{and} \quad Z \perp\!\!\!\perp Y_{x=1,z} \qquad \text{for z=0,1.} \tag{8}$$

(Note that (7) is implied by (2), though they are not equivalent.) This follows from the following simple argument. For $i, j, k \in \{0, 1\}$,

$$\begin{aligned} p(Y_{x=i,z=k} = j) &= p(Y_{x=i,z=k} = j \mid Z = k) \\ &= p(Y_{x=i,z=k} = j, X = i \mid Z = k) \\ &\qquad + p(Y_{x=i,z=k} = j, X = 1 - i \mid Z = k) \\ &\leq p(Y = j, X = i \mid Z = k) + p(X = 1 - i \mid Z = k), \end{aligned} \tag{9}$$

where the first equality follows from (8). It follows that:

$$\max_k p(Y = 1, X = i \mid Z = k) \leq p(Y_{x=i,z=k} = 1) \leq \min_{k^*} 1 - p(Y = 0, X = i \mid Z = k^*),$$

where the lower bound is obtained from (9) taking $j = 0$. The requirement that the lower bound be less than the upper bound, together with (7) then directly implies (4) with $i = 0$, and (5) with $i = 1$. Thus, (4) and (5) will hold in contexts where $Z \not\perp\!\!\!\perp X_z$, for example, where there is confounding between Z and X, or even where

X_z is not well-defined, provided that (7) and (8) hold. Further, we do not require joint independence:

$$Z \perp\!\!\!\perp Y_{x_0}, Y_{x_1}. \tag{10}$$

However, we know of few realistic contexts where one would know that (8) holds, but (10) does not.

Robins (1989) considered the ACE($X \to Y$) under the model given by (7) and (8), deriving what Pearl (2000) calls the "natural bounds"; see also Manski (1990). Although the natural bounds are sharp under the assumption of (7) and (8) alone, they are not sharp under the stronger assumption (1) and (7), for which Pearl derived the bounds. This is interesting given that, as we have seen, both of these independence assumptions, combined with (7), lead to the same set of distributions $p(x, y \mid z)$ for the observables, characterized by (4) and (5). Finally, we note that in fact the ACE bounds derived by Pearl assuming (7) and (1) are also implied by the weaker assumption (10), *i.e.*, without requiring $Z \perp\!\!\!\perp X_{z=0}, X_{z=1}$ (Richardson and Robins, 2011). This appears to contradict a remark in Pearl (2009, p. 395).

4.3. *Inequalities Defining Models without Defiers*

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

$$p(x_1 \mid z_1) \geq p(x_1 \mid z_0), \tag{11}$$

since the left and right sides are the proportions of (AT or CO) and AT respectively. Thus (11) characterizes the observed distributions resulting from Mon_X alone.

Results of Balke and Pearl (1997) imply that the model assuming Mon_X + Ex_{NT} + Ex_{AT} implies the following inequalities:

$$p(y_1, x_0 \mid z_1) \ \leq\ p(y_1, x_0 \mid z_0), \quad p(y_0, x_0 \mid z_1) \ \leq\ p(y_0, x_0 \mid z_0), \tag{12}$$

$$p(y_1, x_1 \mid z_1) \ \geq\ p(y_1, x_1 \mid z_0), \quad p(y_0, x_1 \mid z_1) \ \geq\ p(y_0, x_1 \mid z_0). \tag{13}$$

The inequalities (12) and (13) imply (11), (4) and (5). A distribution $p(y, x \mid z)$ may violate all of the inequalities (12) and (13) simultaneously. However, if (11) holds, then at most one inequality in each of the pairs (12) and (13) may be violated. The inequalities (12) and (13) are invariant to relabeling Y, and to relabeling X and Z simultaneously, but not individually; this is not surprising since relabeling X or Z alone will turn Defiers into Compliers and viceversa.

It may be shown that (12) and (13) characterize the set of distributions $p(y, x|z)$ arising from the potential outcomes model Mon_X + Ex_{NT} + Ex_{AT}. Likewise, the model imposing Mon_X + Ex_{NT} is characterized by (11) and (12), while Mon_X + Ex_{AT} is given by (11) and (13).

An interpretation of (12) and (13) is given by the following lower bound on π_{DE} in the model that imposes Ex_{NT} + Ex_{AT} (but not Mon_X):

$$\pi_{\text{DE}} \geq \max \left\{ \begin{array}{ll} 0, & p(x_1 \mid z_0) - p(x_1 \mid z_1), \\ p(y_1, x_0|z_1) - p(y_1, x_0|z_0), & p(y_0, x_0|z_1) - p(y_0, x_0|z_0), \\ p(y_1, x_1|z_0) - p(y_1, x_1|z_1), & p(y_0, x_1|z_0) - p(y_0, x_1|z_1) \end{array} \right\}; \tag{14}$$

see Richardson and Robins (2010). Requiring that the lower bound be zero, as required by Mon_X, leads directly to the inequalities (11), (12) and (13).

Another interpretation of (12) and (13) 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:

$$\begin{aligned} p(y_1|x_0, z_1) - \min\left(p(y_1, x_0|z_0)/p(x_0|z_1), 1\right) \leq \mathrm{ACDE}_{\mathrm{NT}}(x_0) \leq \\ p(y_1|x_0, z_1) - \max\left(0, 1 - (p(y_0, x_0|z_0)/p(x_0|z_1))\right), \end{aligned} \tag{15}$$

$$\begin{aligned} \max\left(0, 1 - (p(y_0, x_1|z_1)/p(x_1|z_0))\right) - p(y_1|x_1, z_0) \leq \mathrm{ACDE}_{\mathrm{AT}}(x_1) \leq \\ \min\left(p(y_1, x_1|z_1)/p(x_1|z_0), 1\right) - p(y_1|x_1, z_0). \end{aligned} \tag{16}$$

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 (15) iff one of the inequalities (12) is violated; likewise for $\mathrm{ACDE}_{\mathrm{AT}}(x_1)$, (16) and (13). Thus, if Mon_X, and hence (11) holds, then, as mentioned above, at most one inequality in each of the pairs (12) and (13) 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, $\mathrm{ACDE}_{\mathrm{NT}}$ and $\mathrm{ACDE}_{\mathrm{AT}}$ may *both* be bounded away from zero. Thus, under the assumption of No Defiers both $\mathrm{Ex}_{\mathrm{NT}}$ and $\mathrm{Ex}_{\mathrm{AT}}$ may be inconsistent with $p(y, x|z)$.

Finally, we note that in the situation where (12) and (13) hold, the natural bounds on $\mathrm{ACE}(X \to Y)$ are sharp (regardless of whether Mon_X holds or Z_x is undefined). should be 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* $p(y, x|z)$*;* (12) *and* (13) *imply* (11).

Model	Assumptions	Constraints on $p(y, x\|z)$
Saturated	Randomization (1)	None
$\mathrm{Ex}_{\mathrm{NT}}$	(1), Exclusion for NT	(4)
$\mathrm{Ex}_{\mathrm{AT}}$	(1), Exclusion for AT	(5)
$\mathrm{Ex}_{\mathrm{AT}} + \mathrm{Ex}_{\mathrm{NT}}$	(1), Exclusion for AT and NT	(4), (5)
Mon_X	(1), No Defiers	(11)
$\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{NT}}$	(1), No Defiers, Exclusion for NT	(11), (12)
$\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{AT}}$	(1), No Defiers, Exclusion for AT	(11), (13)
Mon_X $+ \mathrm{Ex}_{\mathrm{NT}} + \mathrm{Ex}_{\mathrm{AT}}$	(1), No Defiers, Exclusion for NT and AT	[(11)], (12), (13),

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.

Table 5: *Summary of Flu vaccine data; originally from McDonald e tal. (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	

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), (11), (12) and (13) are violated. For a model without covariates, these probabilities are shown in the first line of Table 7. The posterior probability that at least one of the inequalities (13) fails to hold is 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 (13) 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:

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

for NT, AT and CO under the models Mon_X, Mon_X+$\mathrm{Ex}_{\mathrm{AT}}$, Mon_X+$\mathrm{Ex}_{\mathrm{NT}}$ and Mon_X+$\mathrm{Ex}_{\mathrm{AT}}$+$\mathrm{Ex}_{\mathrm{NT}}$. Under additional exclusion assumptions for compliers, $\gamma^{00}_{\mathrm{CO}} = \gamma^{01}_{\mathrm{CO}}$ and $\gamma^{10}_{\mathrm{CO}} = \gamma^{11}_{\mathrm{CO}}$, $\mathrm{ITT}_{\mathrm{CO}}$ is equal to the Complier Average Causal Effect of X on Y, $\mathrm{ACE}_{\mathrm{CO}}(X \to Y) \equiv E[Y_{X_1} - Y_{X_0} \mid \mathfrak{t}_X] = \gamma^{1\cdot}_{\mathrm{CO}} - \gamma^{0\cdot}_{\mathrm{CO}}$.

In Figure 7, we display the joint posterior distributions over upper and lower bounds on $\mathrm{ITT}_{\mathrm{CO}}$ 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 Sections 2 and 3 of Richardson and Robins (2010).

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, $\mathrm{Ex}_{\mathrm{AT}}$, and $\mathrm{Ex}_{\mathrm{NT}}$ 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.

5.1. *Parametrization of Models with Defiers*

For each v, consider the set of distributions $p(y,x|z,v)$ that result from models assuming both $\mathrm{Ex}_{\mathrm{AT}}$ and $\mathrm{Ex}_{\mathrm{NT}}$, and hence satisfy the inequalities (4) and (5) for each v. In the following development, all models and probability statements are conditional on v, which we suppress in the notation.

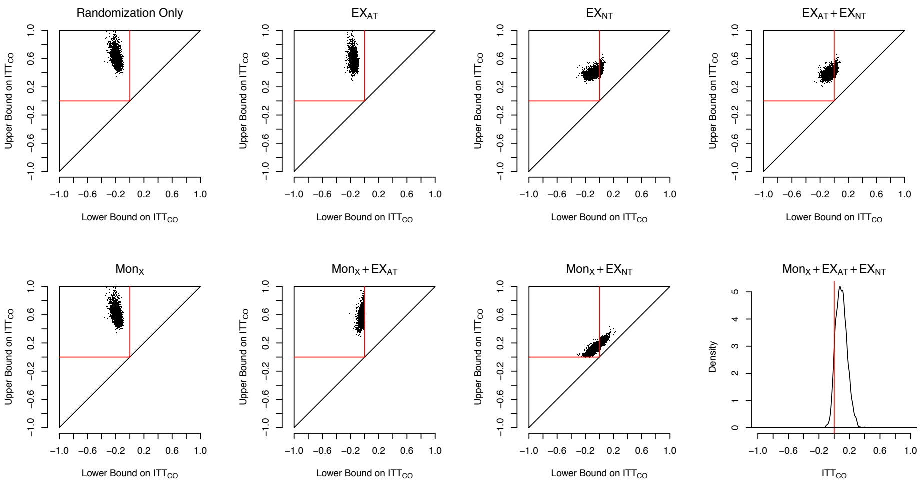


Figure 7: *Posterior distributions for upper and lower bounds on* $\mathrm{ITT}_{\mathrm{CO}}$*; under* Mon_X+$\mathrm{Ex}_{\mathrm{AT}}$+$\mathrm{Ex}_{\mathrm{NT}}$ *the parameter is identified.*

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:

$$\left\{p(y,x|z) \;\middle|\; p(y,x|z_0) \in \Delta_3,\;\; p(y,x|z_1) \in \Delta_3 \cap \bigcap_{i,j\in\{0,1\}} H_{ij}(p(y_{1-i},x_j|z_0))\right\} \quad (17)$$

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 (17) via the parameters: $p(x_1|z_i)$, $p(y_1|x_i,z_0)$ $i=0,1$ and two further parameters, ψ_0, ψ_1 where

$$\psi_i \;\equiv\; \log\left(\frac{p(y_0,x_i|z_1)(1-p(y_1,x_i|z_1)-\{p(y_0,x_i|z_0)\})}{p(y_1,x_i|z_1)(1-p(y_0,x_i|z_1)-\{p(y_1,x_i|z_0)\})}\right). \quad (18)$$

Thus, ψ_i replaces the parameter $p(y_1|x_i,z_1)$. Under (4) and (5), $p(y_1|x_i,z_1)$ is not variation independent of $p(x_1|z_i)$ and $p(y_1|x_i,z_0)$. In contrast, $p(x_1|z_i)$ and $p(y_1|x_i,z_0)$, ψ_0 and ψ_1 are variation independent. The inverse map from the variation independent parameters to $p(y_1|x_i,z_1)$ is given by:

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

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

$$\tilde{\psi}_i \;\equiv\; \log\left(\frac{p(y_0,x_i|z_1)(1-p(y_1,x_i|z_1))}{p(y_1,x_i|z_1)(1-p(y_0,x_i|z_1))}\right), \quad (19)$$

the parameter defined by removing the terms in braces from (18), 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 those for $\mathrm{Ex}_{\mathrm{AT}}$ + $\mathrm{Ex}_{\mathrm{NT}}$.

5.2. *Parametrization of Models without Defiers*

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

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

The model 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:

$$\phi_0 \;\equiv\; \log\left(\frac{p(y_0,x_0|z_1)(1-p(y_1,x_0|z_1)-\{1-p(y_1,x_0|z_0)\})}{p(y_1,x_0|z_1)(1-p(y_0,x_0|z_1)-\{1-p(y_0,x_0|z_0)\})}\right),$$

$$\varphi_1 \;\equiv\; \log\left(\frac{(1-p(y_1,x_1|z_1))(p(y_0,x_1|z_1)-\{p(y_0,x_1|z_0)\})}{(1-p(y_0,x_1|z_1))(p(y_1,x_1|z_1)-\{p(y_1,x_1|z_0)\})}\right).$$

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

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

Hirano *et al.* (2000) give an alternative variation independent parametrization for the observed data distribution under this model. The fact that the Hirano *et al.* model parametrizes the observed data distribution is a consequence of the fact that this model is nonparametrically identified; see also our rejoinder for further discussion.

Mon_X + $\mathrm{Ex}_{\mathrm{AT}}$ 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.

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

Note that the parameters $p(y,x|z_0)$, $\nu_{x|z_1}$, $\tilde{\phi}_0$ and $\tilde{\varphi}_1$ provide an alternative parametrization of the model Mon_X.

Table 6: *Parametrization of Models. Distributions appearing in the parameter list are unrestricted.*

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

5.3. *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 7 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.

Table 7: *Posterior probabilities that inequalities are violated under models that do not impose constraints. The two models without Age used a uniform prior on* $\Delta_3 \times \Delta_3$*; the model with Age used logistic regressions with Normal priors. Columns* (4)*,* (5)*,* (12) *and* (13) *give the probability that at least one inequality is violated;* (12)+(13) *is the probability of at least one violation in both pairs; (12)b is the probability that both inequalities are violated; similarly for (13)b.*

age	copd	(4)	(5)	(11)	(12)	(13)	(12)+(13)	(12) b	(13) b
-	-	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

To illustrate our parametrization, we fitted the four models that include Mon_X. Figure 8 shows posterior distributions on $\mathrm{ITT}_{\mathrm{CO}}$ under $\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{NT}} + \mathrm{Ex}_{\mathrm{AT}}$ in which this parameter is identified, and posterior distributions on bounds under the other three models. Each model was specified via logistic regressions for $p(y|x_0, z_0)$, $p(y|x_1, z_0)$, $p(x|z_0)$ and linear models for $\nu_{x|z_1}$, ϕ_0 (or $\tilde{\phi}_0$) and φ_1 (or $\tilde{\varphi}_1$), again each with intercept, Age, COPD and COPD×Age. Independent $N(0, 3)$ priors were used for all 6×4 coefficients. Sampling was performed via a Metropolis algorithm. The proposal for each of the six GLMs was multivariate normal, mean 0, covariance

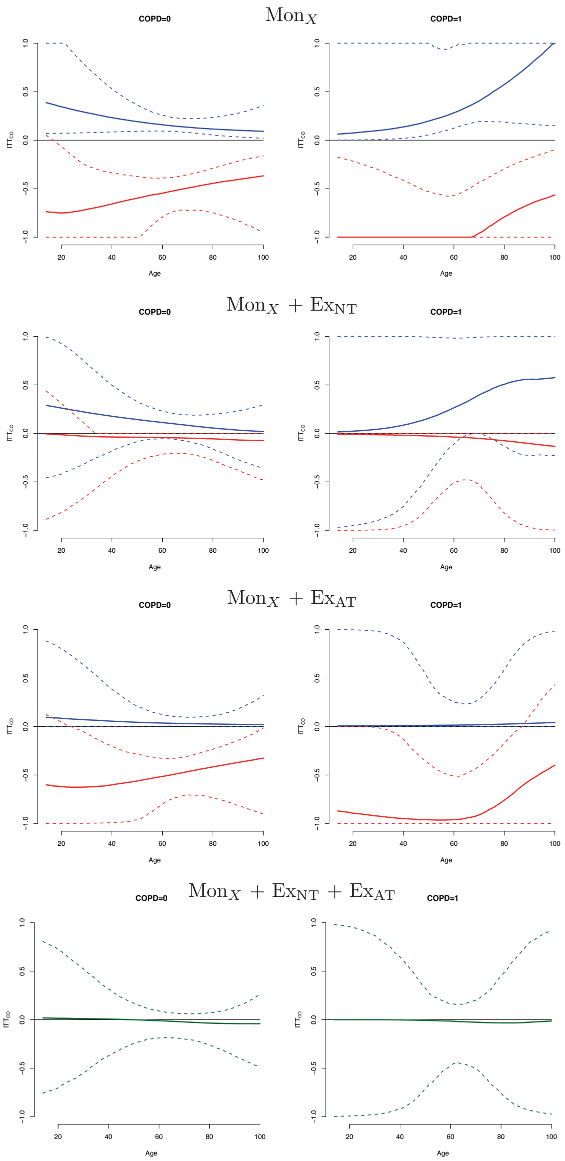


Figure 8: *Flu vaccine data. Posteriors on lower and upper bounds for* $\mathrm{ITT}_{\mathrm{CO}}$*, when not identified, and on* $\mathrm{ITT}_{\mathrm{CO}}$ *itself, when identified, as a function of Age and COPD for the four models which preclude Defiers; medians are solid, pointwise credible intervals are dashed.* $p(y, x|z_0, v)$ *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 *(or* $\tilde{\phi}_0$*) and* φ_1 *(or* $\tilde{\varphi}_1$*) are modelled as linear functions of Age and COPD.*

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. Our results are generally consistent with those obtained without including covariates; see the second row of Figure 8.

6. PROTECTING THE CAUSAL NULL HYPOTHESIS FROM POSSIBLE MODEL MIS-SPECIFICATION

In this section, we consider a successfully randomized clinical trial (RCT) with data available on a vector V of baseline covariates that includes continuous covariates, such as height and weight, and with randomization probabilities $p(Z=1 \mid V)$ that do not depend on V. We further assume that Eqs. (1) and (2) hold for each v; hence, the causal model Ex_{AT} + Ex_{NT} holds within strata defined by V. The induced model for the observed data is characterized by Eqs. (4) and (5) holding for each v.

Under this model, the sharp null hypothesis

$$Y_{x=1} = Y_{x=0} = Y \tag{20}$$

of no causal effect of X on Y for each subject implies both the conditional and unconditional intention to treat (ITT) null hypotheses

$$p(Y = 1 \mid Z = 1, V) - p(Y = 1 \mid Z = 0, V) = 0 \tag{21}$$

and

$$p(Y = 1 \mid Z = 1) - p(Y = 1 \mid Z = 0) = 0. \tag{22}$$

Thus, a test of either of these ITT null hypotheses is a test of the sharp null (20). Since the conditional ITT null (21) implies the unconditional null (22) but not vice versa, a test of the conditional null is preferable. Furthermore, tests that use data on V may be more powerful than tests that ignore V.

Although the cost of RCTs is often an order of magnitude greater than the cost of an observational study, the U.S. FDA will generally only license a new drug if benefit has been demonstrated in such a randomized trial. The primary reason behind this policy is that, in contrast to observational studies, the sharp null can be empirically tested when Eq. (2) holds. Thus, it is critical to analyze these trials with a robust methodology that guarantees that, under the ITT null, the estimator of $p(Y = 1 \mid Z = 1, V) - p(Y = 1 \mid Z = 0, V)$ converges to zero in probability under the true distribution of (Z, X, Y, V) as the sample size $n \to \infty$, even under model mis-specification.

Unfortunately, the procedure used to estimate the joint distribution of (X, Y) given (Z, V) in the previous section does not fulfill this guarantee. Specifically, suppose we specify a variation independent parametric model for $p(x|z_1, v)$, $p(y, x \mid z_0, v)$, $\psi_0(v), \psi_1(v)$ and a smooth prior for its parameters. If, as will essentially always be the case in practice, these parametric models are mis-specified, then the posterior distribution of the function

$$\text{ITT}(v) = p(Y = 1 \mid Z = 1, V = v) - p(Y = 1 \mid Z = 0, V = v)$$

will generally concentrate on a non-zero function under the ITT null that $\text{ITT}(v) = 0$ for almost all v. This follows from the fact that, in large trials, the posterior distribution of $\text{ITT}(v)$ will be centered on the MLE of $\text{ITT}(v)$ and the MLE is generally inconsistent under mis-specification. Thus, in large trials, we will falsely reject the sharp null hypothesis even when true. As a consequence, not only has the large sum spent on the trial been wasted but, more importantly, a drug without benefit may become licensed.

Example: As a concrete illustration of the danger of mis-specification, we generated 5000 samples from the following data generating process, in which V is a covariate taking three states:

$$\begin{aligned} Z &\sim \text{Ber}(1/2), \\ V &\sim \text{Uniform}(\{0,1,2\}), \\ Y_{x=0,z=0} = Y_{x=0,z=1} = Y_{x=1,z=0} = Y_{x=1,z=1} &\sim \text{Ber}(1/2), \end{aligned}$$

and

$$X_z \mid Y_{x,z} = y, V = v \;\sim\; \text{Ber}\Big(\text{expit}(-4+y+z+\frac{5}{2}v-\frac{3}{2}v^2+y\cdot(z+v)+v\cdot z(1+2y))\Big).$$

Note that it follows from this scheme that $Y \perp\!\!\!\perp \{Z, V\}$, so that the conditional ITT null hypothesis clearly holds. We fitted linear models (in V) for the six parameters $\text{logit}(p(x_1 \mid z_i))$, $\text{logit}(p(y_1 \mid x_i, z_0))$ and ψ_i, $i \in \{0,1\}$ for $\text{Ex}_{\text{AT}} + \text{Ex}_{\text{NT}}$; see Eq. (18).

We performed inference via MCMC with a burn-in of 1000, retaining 5000 iterations. The posterior distributions for each of the ITT effects,

$$p(y = 1|z = 1, v) - p(y = 1|z = 0, v)$$

are shown in Table 8. As can be seen, the posterior distribution indicates ITT effects in two strata, even though none was present in the generating process.

Theorems 1 and 2 in the appendix provide necessary and sufficient conditions for a parametric model to be robust to mis-specification under the ITT null. These theorems include as special cases the results of Rosenblum and Van Der Laan (2009).

Table 8: *Summary of posterior for ITT(v) under ITT null with mis-specification*

v	2.5%	Mean	97.5%
0	−0.02289	0.02041	0.06190
1	−0.12820	−0.09863	−0.07088
2	0.02395	0.06915	0.11590

6.1. *ITT-Null-Robust Parametrization of* $\text{Ex}_{\text{AT}} + \text{Ex}_{\text{NT}}$

The key to constructing a parametric Bayes estimator robust to mis-specification under the ITT null is to parametrize, for each v, the set of distributions (17) for the observed data corresponding to a model constrained by (4) and (5) as follows: for each $V = v$, we have the following six variation independent parameters:

$$\pi_i \equiv p(y_1 \mid z_i), \quad p(x_1 \mid y_i, z_0), \quad \zeta_i, \qquad \text{for } i \in \{0,1\},$$

where:

$$\begin{aligned}
\zeta_0 \quad \equiv \quad & \frac{(1-p_X^{01})}{p_X^{01}} \frac{[1 - p_X^{10}\pi_0 - p_X^{01}(1-\pi_1)]}{[1-(1-p_X^{10})\pi_0 - (1-p_X^{01})(1-\pi_1)]} \\
& \times \left(1 + \frac{(p_X^{10} - p_X^{01})(1-\pi_0)}{(1-p_X^{10})}\right)^{-p_X^{10}/(p_X^{10}-p_X^{01})} \left(1 - \frac{(p_X^{10} - p_X^{01})(1-\pi_0)}{p_X^{10}}\right)^{-(1-p_X^{10})/(p_X^{10}-p_X^{01})} \\
& \times \left(1 + \frac{(p_X^{10} - p_X^{01})(1-\pi_1)}{(1-p_X^{10})}\right)^{p_X^{01}/(p_X^{10}-p_X^{01})} \left(1 - \frac{(p_X^{10} - p_X^{01})(1-\pi_1)}{p_X^{10}}\right)^{(1-p_X^{01})/(p_X^{10}-p_X^{01})},
\end{aligned}$$

$$\begin{aligned}
\zeta_1 \quad \equiv \quad & \frac{(1-p_X^{11})}{p_X^{11}} \frac{[1 - p_X^{00}(1-\pi_0) - p_X^{11}\pi_1]}{[1-(1-p_X^{00})(1-\pi_0) - (1-p_X^{11})\pi_1]} \\
& \times \left(1 + \frac{(p_X^{00} - p_X^{11})\pi_0}{(1-p_X^{00})}\right)^{-p_X^{00}/(p_X^{00}-p_X^{11})} \left(1 - \frac{(p_X^{00} - p_X^{11})\pi_0}{p_X^{00}}\right)^{-(1-p_X^{00})/(p_X^{00}-p_X^{11})} \\
& \times \left(1 + \frac{(p_X^{00} - p_X^{11})\pi_1}{(1-p_X^{00})}\right)^{p_X^{11}/(p_X^{00}-p_X^{11})} \left(1 - \frac{(p_X^{00} - p_X^{11})\pi_1}{p_X^{00}}\right)^{(1-p_X^{11})/(p_X^{00}-p_X^{11})},
\end{aligned}$$

and $p_X^{ij} \equiv p(x_1 \,|\, y_i, z_j)$. Thus, we have replaced $p(x_1 \,|\, y_i, z_1)$ of the standard variation dependent parametrization of this model by ζ_i, $i = 0, 1$. Unlike our previous parametrizations, the inverse map from the variation independent parameters to $p(x_1 \mid y_i, z_1)$, $i = 0, 1$, is not available in closed form.

To model $p(x, y, z, v)$ when V contains continuous covariates, we specify parametric models $\zeta_i(v; \alpha_0)$, $p(x_1 \mid y_i, z_0, v; \alpha_1)$ $i = 0, 1$ and $f(v; \alpha_2)$ as well as a logistic regression model $\mathrm{expit}(Zm(V; \tau) + q(V; \alpha_3))$ for $p(y_1|z, v)$ satisfying $m(V; \tau) = 0$ if and only if τ is the zero vector; the parameter vectors α_0, α_1, α_2, α_3, τ are variation independent. Thus, under this model, $\tau = 0$ if and only if the conditional ITT null hypothesis $Y \perp\!\!\!\perp Z \mid V$ holds.

In the appendix, we show, following Rosenblum and Van Der Laan (2009), that if we choose $q(V; \alpha_3) = \alpha_3^T q^*(V)$ such that each component of $\partial m(V; 0)/\partial\tau$ is in the linear span of the components of $q^*(V)$, then, under the conditional ITT null, the MLE of τ converges to its true value of 0 at rate $n^{-1/2}$ even if the models $\zeta_i(v; \alpha_0)$, $p(x_1|y, z_0, v; \alpha_1)$, $f(v; \alpha_2)$ and $p(y_1|z_0, v; \alpha_3) = \mathrm{expit}(\alpha_3^T q^*(v))$ are all mis-specified. As a consequence, under a smooth prior $p(\tau, \alpha)$, the posterior distribution of τ and of $\mathrm{ITT}(v)$ will concentrate on the zero function in large samples.

Note that we can always guarantee the linear span condition holds by choosing $q^*(V)$ to include $\partial m(V; 0)/\partial\tau$ as a subvector.

6.2. *ITT-Null-Robust Parametrization of* $\mathrm{Mon}_X + \mathrm{Ex}_{\mathrm{AT}} + \mathrm{Ex}_{\mathrm{NT}}$

To obtain a parametric Bayes estimator with the aforementioned robustness property under the conditional ITT null for the model that excludes Defiers, we parametrize, for each v, the set of distributions for the observed data constrained by (12) and (13) as follows: for each $V = v$, we have the following six variation independent parameters:

$$\begin{aligned}
\pi_i &\equiv p(y_1 \mid z_i), \\
\xi_i &\equiv \frac{p_X^{i1}}{(1-p_X^{i1})}\left[\pi_0 - (1-p_X^{i1})\pi_1\right] \times \pi_0^{-1/p_X^{i1}} \times \pi_1^{(1-p_X^{i1})/p_X^{i1}}, \\
\kappa_i &\equiv \frac{p_X^{i0}\pi_0^2}{[(1-p_X^{i0})\pi_0 - (1-p_X^{i1})\pi_1][p_X^{i1}\pi_1 - p_X^{i0}\pi_0]} \times \left(\frac{\pi_1}{\pi_0}\right)^{(2p_X^{i1}-1)/(p_X^{i1}-p_X^{i0})},
\end{aligned}$$

for $i \in \{0,1\}$, where $p_X^{ij} \equiv p(x_1 \mid y_i, z_j)$.

In the appendix, we prove that any parametric submodel with $p(y_1 \mid z_i)$ modelled as in the previous subsection will enjoy the same robustness properties under the conditional ITT null, even if the models for $\xi_i(v;\alpha_0)$, $\kappa_i(v;\alpha_1)$, $f(v;\alpha_2)$ and $p(y_1|z_0,v;\alpha_3) = \text{expit}(\alpha_3^T q^*(v))$ are all mis-specified.

Though we do not do so here, our approach may be extended to the other six potential outcome models in which fewer exclusion restrictions hold.

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 Section 3, and to Sander Greenland and Adrian Raftery for helpful suggestions.

REFERENCES

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92**, 1171–1176.

Bickel, P. J., Klaasen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: John Hopkins University Press.

Bonet, B. (2001). Instrumentality tests revisited. *Proc. 17th Conf. on Uncertainty in Artificial Intelligence*, 48–55.

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

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

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

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

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

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

Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research* **3**, 405–430.

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

Hudgens, M. G., Hoering, A. and Self, S.G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22**(14), 2281–2298.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statist. Probab. Lett.* **78**, 144–149.

Imbens, G. and Rubin, D. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305–327.

Kaufman, S., Kaufman, J.S. and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *J. Statist. Planning and Inference* **139**, 3473–3487.

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

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

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

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **X**, 1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Statist. Science* **5** 463–472, 1990.

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

Pearl, J. (2009). *Causality* (2nd edition). Cambridge: 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., Evans, R.J. and Robins, J. M. (2011). Variation independent parameterizations for convex models. *Tech. Rep.*, University of Washington, USA.

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

Richardson, T. S. and Robins, J. M. (2011). Discrete instrumental variable models for dichotomous outcomes. *Tech. Rep.*, University of Washington, USA.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Applications to control of the healthy worker survivor effect. *Mathematical Modeling* **7**, 1393–1512.

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

Robins, J. M., Richardson, T. S. and Rotnitzky, A. (2011). Robustness of parametric submodels. *Tech. Rep.*, Harvard School of Public Health, USA.

Robins, J. M., Richardson, T. S. and Spirtes, P. (2009). Identification and inference for direct effects. *Tech. Rep.*, University of Washington, USA.

Rosenblum, M. and Van Der Laan, M. J. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* **65**, 937–945.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educational Psychology* **66**, 688–701.

Rubin, D .B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine* **17**, 371–385.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian J. Statist.* **3**, 161–170.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educational and Behavioral Statistics* **28**, 353–368.

APPENDIX

To prove the claims in Section 6 we establish some general properties of parametric models under mis-specification. For this purpose, we require some additional semiparametric concepts; see Bickel *et al.* (1993).

Given a model M and smooth functional $\tau : F \mapsto \tau(F)$ mapping $F \in M$ into $\mathbb{R}^d$, we wish to estimate $\tau(F)$ based on n i.i.d. observations O_i; $i = 1, \ldots, n$, from an unknown distribution $F \in M$. In this appendix, our goal is to determine the high level conditions required for the MLE of τ based on a given parametric submodel of M to remain consistent for τ at certain laws $F \in M$ that lie outside the submodel.

Formally, a parametric submodel of M is the range $R(h) = \{F_{h,\alpha}; \alpha \in A\} \subset M$ of a smooth map $h : \alpha \mapsto F_{h,\alpha}$ with domain an open set $A \in \mathbb{R}^l$ (Bickel *et al.*, 1993, p. 13). Write $f_h\{o;\alpha\}$ for the density of $F_{h,\alpha}$ with respect to a common dominating measure μ. For convenience, we henceforth suppress the dependence on h and write $f\{o;\alpha\}$ and F_α. Let $\widetilde{\tau}(\alpha) = \tau(F_\alpha)$ be the value of τ at F_α.

Let $M(b) = \{F \in M; \tau(F) = b\}$ be the submodel of M on which the functional τ takes the value $b \in \mathbb{R}^d$. Let $R(h,b) = \{F_\alpha; \alpha \in A, \widetilde{\tau}(\alpha) = b\}$ be the submodel of the parametric model $R(h) = \{F_\alpha; \alpha \in A\}$ contained in $M(b)$. Let $A(b) = \{\alpha \in A; \widetilde{\tau}(\alpha) = b\}$ be the pre-image of $R(h,b)$. Let

$$\begin{aligned} \widehat{\alpha} &= \arg\max_{\alpha \in A} \mathbb{P}_n \left[\log\left\{f\left[O;\alpha\right]\right\}\right] \text{ and} \\ \alpha(F) &= \arg\max_{\alpha \in A} E_F \left[\log\left\{f\left[O;\alpha\right]\right\}\right] \end{aligned}$$

be maximizers of the empirical and expected log-likelihood, where $\mathbb{P}_n$ denotes a sample average. Similarly, let

$$\begin{aligned} \widehat{\alpha}(b) &= \arg\max_{\alpha \in A(b)} \mathbb{P}_n \left[\log\left\{f\left[O;\alpha\right]\right\}\right] \text{ and} \\ \alpha(b;F) &= \arg\max_{\alpha \in A(b)} E_F \left[\log\left\{f\left[O;\alpha\right]\right\}\right] \end{aligned}$$

be the maximizers of the empirical and expected "profile" log-likelihood given $\widetilde{\tau}(\alpha) = b$. Thus, $\alpha(b;F)$ is the maximizer of $E_F\left[\log\left\{f\left[O;\alpha\right]\right\}\right]$ over all $\alpha \in A$ subject to the d constraints $\widetilde{\tau}(\alpha) = b$. Note $\alpha\left\{F_{\alpha^*}\right\} = \alpha^*$, as the expected log likelihood is maximized at the true density. Under regularity conditions $\widehat{\alpha}$ and $\widehat{\alpha}(b)$, respectively, converge to $\alpha(F)$ and $\alpha(b;F)$ at rate $n^{-1/2}$.

Let $S(\alpha) = \partial \log\{f[O;\alpha]\}/\partial\alpha$ be the score for α evaluated at F_α. Typically $\alpha(F)$ is the (assumed) unique solution to $E_F[S(\alpha)] = 0$ for all $F \in M$.

Since $E_{F_\alpha}[S(\alpha)S(\alpha)^T]^{-1}S(\alpha)$ is the influence function for α, taking derivatives we obtain that $IF_{\tau,par}(\alpha) = \{\partial\widetilde{\tau}(\alpha)/\partial\alpha\}^T E_{F_\alpha}[S(\alpha)S(\alpha)^T]^{-1}S(\alpha)$ is the d-dimensional "parametric" efficient influence function for τ in the parametric model $R(h) = \{F_\alpha; \alpha \in A\}$ at F_α. Further, $\text{var}_{F_\alpha}\{IF_{\tau,par}(\alpha)\}$ is the parametric Cramér–Rao variance bound for τ at F_α

For $F_\alpha \in M(b)$, let $\Lambda(\alpha)$ and $\Lambda(b;\alpha)$ be the closed linear span in $L_2(F_\alpha)$ of the scores for all parametric submodels in M and $M(b)$, respectively, that include F_α. By definition, $\Lambda(\alpha)$ and $\Lambda(b;\alpha)$ are the tangent spaces for models M and $M(b)$ at F_α, and $\Lambda(b;\alpha)$ is the nuisance tangent space for τ in model M at F_α. Further, the efficient influence function $IF_\tau(F_\alpha)$ for τ in model M at F_α is the unique element of $\Lambda(\alpha)$ satisfying $E_{F_\alpha}\left[IF_\tau(F_\alpha)D\right] = \partial\tau\left(F(t)\right)/\partial t|_{t=0}$ for all $D \in \Lambda(\alpha)$ and $F(t) \in M$ a parametric model with parameter t with $F(0) = F_\alpha$ and score at F_α equal to D. In particular, $E_{F_\alpha}\left[IF_\tau(F_\alpha)D\right] = 0$ if $D \in \Lambda(b;\alpha)$. Then $\text{var}_{F_\alpha}\{IF_\tau(F_\alpha)\}$ is the semiparametric Cramér–Rao variance bound for τ in model M at F_α. Note $\text{var}_{F_\alpha}\{IF_{\tau,par}(\alpha)\} \leq \text{var}_{F_\alpha}\{IF_\tau(F_\alpha)\}$ with equality if and only if $IF_{\tau,par}(\alpha) = IF_\tau(F_\alpha)$ w.p.1.

Robins *et al.* (2011) prove the following two Theorems.

Theorem 1 *Given a fixed* b*, suppose for all* $F \in M(b)$*,* $E_F\left[S(\alpha)\right] = 0$ *has a unique solution. If* $E_F[S(\alpha(b;F))] = 0$ *for all* $F \in M(b)$*, then, for all* $\alpha \in A(b)$*,*

$$IF_\tau(F_\alpha) = IF_{\tau,par}(\alpha).$$

Note the conclusion of Theorem 1 can also be expressed as equality of the parametric and semiparametric (for model M) Cramér–Rao variance bounds for τ at $F_\alpha, \alpha \in A(b)$.

Corollary 1.1 *Suppose for all* $F \in M(b)$*,* $E_F[S(\alpha)] = 0$ *has a unique solution. Then the following hold:*

(i) *If* $IF_\tau(F_\alpha) \neq IF_{\tau,par}(\alpha)$ *for some* $F_\alpha \in M(b)$*, then there exists* $F^* \in M(b)$ *such that* $\widetilde{\tau}\{\alpha(F^*)\} \neq b$*.*

(ii) *The MLE* $\widehat{\tau} \equiv \widetilde{\tau}(\widehat{\alpha})$ *is not consistent for* $\tau = b$ *at* F^**.*

Proof: (i) By Theorem 1 there exists $F^* \in M(b)$ such that $E_{F^*}\left[S(\alpha(b;F^*))\right] \neq 0$. Hence $\widetilde{\tau}\{\alpha(F^*)\}$ cannot equal $\widetilde{\tau}\{\alpha(b;F^*)\} \equiv b$.

(ii) Since $\widetilde{\tau}\{\alpha(F^*)\}$ is the limit of the MLE $\widetilde{\tau}\{\widehat{\alpha}\}$, (ii) follows. □

Remark: Robins *et al.* (2011) prove the following stronger result. Let

$$Q = \{\alpha; \alpha \in A\left(b\right) \text{ and } IF_\tau(F_\alpha) \neq IF_{\tau,par}(\alpha)\}.$$

Then there exists an injective map from $Q \rightarrow M(b)$ such that $\widetilde{\tau}\{\alpha(F)\} \neq b$ for all F in the range of the map.

Theorem 2 *Suppose for all* $F \in M(b)$*,* $E_F[S(\alpha)] = 0$ *has a unique solution. Then if* (i) *for all* $\alpha \in A(b)$*, the parametric and semiparametric influence functions are equal, i.e.,* $IF_\tau(F_\alpha) = IF_{\tau,par}(\alpha)$ *w.p. 1 for all* $\alpha \in A(b)$*, and* (ii) *for all* $F \in M(b)$*,* $(f(O)/f(O;\alpha(b;F))) - 1$ *is contained in the tangent space* $\Lambda(b;\alpha(b;F))$ *of* $M(b)$ *at* $\alpha(b;F)$*, then*

$$E_F\left[S(\alpha(b;F))\right] = 0 \textit{ for all } F \in M(b). \tag{23}$$

Remark: A sufficient condition for (ii) is that the model $M(b)$ is convex: that is, if F_1 and F_2 are in $M(b)$ then so is the law $\lambda F_1 + (1-\lambda)F_2$ for $\lambda \in [0,1]$.

Corollary 2.1 *Under the hypotheses of Theorem 2, for all* $F \in M(b)$*,*

(i) $\widetilde{\tau}\left(\alpha\left(F\right)\right) = b$*;*

(ii) *under regularity conditions, the parametric MLE* $\widehat{\tau}$ *is a consistent, asymptotically normal (CAN) estimator of* $\tau(F) = b$*.*

Proof: (i) The conclusion of Theorem 2 implies that $\widetilde{\tau}\left(\alpha\left(b;F\right)\right) = \widetilde{\tau}\left(\alpha\left(F\right)\right)$ for all $F \in M(b)$. But $\widetilde{\tau}\left(\alpha\left(b;F\right)\right) \equiv b$ and $\tau\left(F\right) \equiv b$. (ii) Under standard regularity conditions, the MLE is a CAN estimator of its limit.

□

6.3. *Proving the Robustness Claims of Section 6*

We first consider the model $\mathrm{Ex}_{\mathrm{AT}} + \mathrm{Ex}_{\mathrm{NT}}$, which allows for Defiers.
Let M be a model for $O = (V, Z, X, Y)$ characterized by:

(i) $p(Z = 1|V) = c$ is a known constant;

(ii) $\mathrm{logit}[p(y_1 \mid z_i, v)] \in \{Zm(v;\tau) + q(v);\ \tau \in \mathbb{R}^d,\ q : v \mapsto q(v)$ unrestricted, $m(\cdot;\cdot)$ known satisfying $m(\cdot;\tau) = 0 \Leftrightarrow \tau = 0\}$;

(iii) $\zeta_i(v)$, $\mathrm{logit}[p(x_1 \mid y_i, z_0, v)]$, $i = 0, 1$ and $f(v)$ are all unrestricted.

It follows from the construction of the parameters ζ_0 and ζ_1 that for each v, Eqs. (4) and (5) will hold with strict inequalities.

Let $M(0)$ be the submodel of M with $\tau = 0$, in which the conditional ITT Null (21) holds. Thus, $M(0)$ is the model characterized by the constraints (4) and (5) and $Y \perp\!\!\!\perp Z \mid V$.

Henceforth, we use the semiparametric notation introduced at the start of the appendix. Let $R(h) = \left\{F_\alpha; \alpha \in A \subseteq \mathbb{R}^l\right\}$ be a parametric submodel $M_{sub,par}$ of M characterized by known functions of $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{R}^l$: $\zeta_i(v; \alpha_0)$, $p(x_1|y_i, z_0, v; \alpha_1)$, $f(v; \alpha_2)$, $q(v; \alpha_3) = \alpha_3^T q^*(v)$, $\alpha_4 = \tau$ with the α_j variation independent and with each component of $\partial m(V; 0)/\partial\tau$ in the linear span of the components of $q^*(V)$.

Such a parametric submodel exists because our ITT-null-robust parametrization is variation independent. We further assume that there is a unique solution to $E_F[S(\alpha)] = 0$, for all $F \in M(0)$.

Our goal is to prove the following:

Theorem 3 *For $F \in M(0), \alpha(F)$ solving $E_F[S(\alpha)] = 0$ satisfies the constraint $\widetilde{\tau}(\alpha(F)) \equiv \alpha_4(F) = 0$,* i.e.*, $\alpha(F) = \alpha(0; F)$.*

Theorem 3 implies that, under regularity conditions, the parametric MLE $\widehat{\tau}$ is a consistent, asymptotically normal (CAN) estimator of $\tau(F) = 0$, for all $F \in M(0)$, *i.e.*, under mis-specification of the parametric (nuisance) submodels.

Remark: A natural approach to a proof is to establish that the premises of Theorem 2 hold and then appeal to its corollary. We shall see that, although premise (i) holds, premise (ii) does not. However, a minor fix gets around this difficulty.

Before proving Theorem 3, we prove the following lemma that establishes premise (i) of Theorem 2.

Lemma 1 *For all $\alpha \in A(0)$, the parametric and semiparametric influence functions are equal, i.e.,*

$$IF_\tau(F_\alpha) = IF_{\tau,par}(\alpha) \text{ w.p. 1 for all } \alpha \in A(0). \tag{24}$$

Furthermore, they depend on α only through α_3 and on the data only through (Z, Y, V). (Recall that $\alpha \in A(0)$ iff $\alpha_4 \equiv \tau = 0$, so α_4 is fixed.)

Proof: Write $S(\alpha) = (S_{\alpha_k}(\alpha); k = 0, 1, 2, 3, 4)$. Consider a particular $F_\alpha \subset M(0)$. Now $IF_{\tau,par}(\alpha) = IF_\tau(F_\alpha)$ if and only if

$$S_{\tau,\mathit{eff}}(\alpha) \equiv S_{\alpha_4} - \Pi_{F_\alpha}\left[S_{\alpha_4}(\alpha)|\Lambda(0;\alpha)\right] = S_{\tau,\mathit{eff,par}}(\alpha) \equiv S_{\alpha_4} - \Pi_\alpha\left[S_{\alpha_4}(\alpha)|S_{\alpha\backslash\alpha_4}(\alpha)\right]$$

where $S_{\alpha\backslash\alpha_4} = (S_{\alpha_0}, \ldots, S_{\alpha_3})$ and Π_{F_α} is the projection operator in $L_2(F_\alpha)$. This follows from the fact that $IF_{\tau,par}(\alpha) = E_{F_\alpha}\left[S_{\tau,\mathit{eff},par}(\alpha)^{\otimes 2}\right]^{-1} S_{\tau,\mathit{eff},par}(\alpha)$ and $IF_\tau(F_\alpha) = E_{F_\alpha}\left[S_{\tau,\mathit{eff}}(\alpha)^{\otimes 2}\right]^{-1} S_{\tau,\mathit{eff}}(\alpha)$.

Now, the likelihood for one observation is

$$\begin{aligned} f(Y,X,V,Z) &= \left\{p_y(Z,V;\alpha_3,\alpha_4)^Y \{1-p_y(Z,V;\alpha_3,\alpha_4)\}^{1-Y}\right\} \\ &\quad \times \left\{p_x^{z=0}(Y,V;\alpha_1)^X \left\{1-p_x^{z=0}(Y,V;\alpha_1)^{1-X}\right\}\right\}^{I(Z=0)} \\ &\quad \times \left\{p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)^X \left\{1-p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)^{1-X}\right\}\right\}^{I(Z=1)} \\ &\quad \times f(V;\alpha_2)c^Z(1-c)^{1-Z}, \end{aligned}$$

where $p_y(z,v;\alpha_3,\alpha_4) = p(y_1 \mid z,v;\alpha_3,\alpha_4)$, $p_x^{z=0}(y,v;\alpha_1) = p(x_1 \mid y,z_0,v;\alpha_1)$, etc. Thus

$$S_{\alpha_0}(\alpha) = I(Z=1)\left\{X - p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}\left\{\frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)/\partial\alpha_0}{p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\left\{1-p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}}\right\},$$

$$\begin{aligned} S_{\alpha_1}(\alpha) &= I(Z=0)\left\{X - p_x^{z=0}(Y,V;\alpha_1)\right\}\frac{\partial p_x^{z=0}(Y,V;\alpha_1)/\partial\alpha_1}{p_x^{z=0}(Y,V;\alpha_1)\left\{1-p_x^{z=0}(Y,V;\alpha_1)\right\}} \\ &\quad + I(Z=1)\left\{X - p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}\frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)/\partial\alpha_1}{p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\left\{1-p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}}, \end{aligned}$$

$$S_{\alpha_2}(\alpha) = \partial\left\{\log f(V;\alpha_2)\right\}/\partial\alpha_2,$$

$$\begin{aligned} S_{\alpha_3}(\alpha) &= \left\{Y - p_y(Z,V;\alpha_3,\alpha_4)\right\}\frac{\partial p_y(Z,V;\alpha_3,\alpha_4)/\partial\alpha_3}{p_y(Z,V;\alpha_3,\alpha_4)\left\{1-p_y(Z,V;\alpha_3,\alpha_4)\right\}} \\ &\quad + I(Z=1)\left\{X - p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}\left\{\frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)/\partial\alpha_3}{p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\left\{1-p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}}\right\}, \end{aligned}$$

$$\begin{aligned} S_{\alpha_4}(\alpha) &= \left\{Y - p_y(Z,V;\alpha_3,\alpha_4)\right\}\frac{\partial p_y(Z,V;\alpha_3,\alpha_4)/\partial\alpha_4}{p_y(Z,V;\alpha_3,\alpha_4)\left\{1-p_y(Z,V;\alpha_3,\alpha_4)\right\}} \\ &\quad + I(Z=1)\left\{X - p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}\left\{\frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)/\partial\alpha_4}{p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\left\{1-p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)\right\}}\right\}. \end{aligned}$$

Richardson *et al.* (2011) prove that under our ITT-null-robust parametrization, when $\alpha_4 = 0$,

$$\frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)}{\partial\alpha_4} = \frac{\partial p_x^{z=1}(Y,V;\alpha\backslash\alpha_2)}{\partial\alpha_3} = 0.$$

Our parametrization was carefully constructed to ensure that these derivatives were zero when $\alpha_4 = 0$. Consequently, under the ITT null ($\alpha_4 = 0$) $S_{\alpha_3}(\alpha)$ and $S_{\alpha_4}(\alpha)$ are functions only of α_3 and the data (Y,Z,V); crucially they are not functions of α_0, α_1, α_2 and X. Let E_α be shorthand for E_{F_α}. At $\alpha_4 = 0$,

$$E_\alpha\left[S_{\alpha_4}(\alpha)\left\{S_{\alpha\backslash(\alpha_3,\alpha_4)}(\alpha)\right\}^T\right] = E_\alpha\left[S_{\alpha_3}(\alpha)\left\{S_{\alpha\backslash(\alpha_3,\alpha_4)}(\alpha)\right\}^T\right] = 0$$

since for $i = 0, 1$, and $j = 3, 4$,

$$E_\alpha \left[S_{\alpha_j}(\alpha) \{ S_{\alpha_i}(\alpha) \}^T \right] = E_\alpha \left[S_{\alpha_j}(\alpha) E_\alpha \left[S_{\alpha_i}(\alpha)^T \middle| Y, V, Z \right] \right] = 0$$

and

$$E_\alpha \left[S_{\alpha_2}(\alpha) S_{\alpha_j}(\alpha)^T \right] = E_\alpha \left\{ S_{\alpha_2}(\alpha) E_\alpha \left[S_{\alpha_j}(\alpha)^T \middle| V \right] \right\} = 0.$$

Thus, $\Pi_{F_\alpha} \left[S_{\alpha_4}(\alpha) \,|\, S_{\alpha \backslash \alpha_4}(\alpha) \right] = \Pi_{F_\alpha} \left[S_{\alpha_4}(\alpha) \,|\, S_{\alpha_3}(\alpha) \right]$. Now at $\alpha_4 = 0$, $p_y(Z, V; \alpha_3, \alpha_4) = E_{\alpha_3}[Y|V]$,

$$\frac{\partial p_y(Z, V; \alpha_3, \alpha_4)/\partial \alpha_4}{p_y(Z, V; \alpha_3, \alpha_4)\{1 - p_y(Z, V; \alpha_3, \alpha_4)\}} \quad = \quad Z \partial m(V; 0)/\partial \alpha_4$$

and

$$\frac{\partial p_y(Z, V; \alpha_3, \alpha_4)/\partial \alpha_3}{p_y(Z, V; \alpha_3, \alpha_4)\{1 - p_y(Z, V; \alpha_3, \alpha_4)\}} = q^*(V).$$

Hence,

$$S_{\alpha_3}(\alpha) = [Y - E_{\alpha_3}(Y|V)]\, q^*(V), \qquad S_{\alpha_4}(\alpha) = \{Y - E_{\alpha_3}[Y|V]\}\, Z \{\partial m(V; 0)/\partial \tau\}.$$

The argument just given above applies to any parametric submodel contained in $M(0)$ and containing F_α. Therefore, when $\alpha_4 = 0$, $\Lambda(0; \alpha) = \Lambda_{(0,1,2)}(0; \alpha) \oplus \Lambda_3(0; \alpha)$ with $\Lambda_{(0,1,2)}(0; \alpha)$ and $\Lambda_3(0; \alpha)$ orthogonal under F_α. Here

$$\Lambda_3(0; \alpha) = \{ (Y - E_{\alpha_3}(Y|V))\, q(V)\, ; q(\cdot) \text{ unrestricted} \}$$

and $\Lambda_{(0,1,2)}(0; \alpha)$ is the linear span of scores corresponding to the set of unrestricted functions and densities $\zeta_i(v), p(x_1|y_i, z_0, v)$, $f(v)$. The argument also implies that the score $S_{\alpha_4}(\alpha) \equiv S_\tau(\alpha)$ for $\tau = \alpha_4$ is orthogonal to $\Lambda_{(0,1,2)}(0; \alpha)$. Thus, when $\alpha_4 = 0$, $S_{\tau, \mathit{eff}}(\alpha) = S_{\alpha_4}(\alpha) - \Pi_{F_\alpha}[S_{\alpha_4}(\alpha) \,|\Lambda(0; \alpha)] = S_\tau(\alpha) - \Pi_{F_\alpha}[S_\tau(\alpha) \,|\Lambda_3(0; \alpha)]$. One can check that $\Pi_{F_\alpha}[S_\tau(\alpha) \,|\Lambda_3(0; \alpha)] = \{Y - E_{\alpha_3}(Y|V)\} \{\partial m(V; 0)/\partial \tau\} c$ with $c = E[Z|V]$. Hence, $S_{\tau, \mathit{eff}}(\alpha) = S_{\tau, \mathit{eff,par}}(\alpha)$ if

$$\{Y - E_{\alpha_3}(Y|V)\} \{\partial m(V; 0)/\partial \tau\} \{E[Z|V]\} = \Pi_{F_\alpha}[S_{\alpha_4}(\alpha) \,|S_{\alpha_3}(\alpha)].$$

But by $\partial m(V; 0)/\partial \tau$ in the span of the components of $q^*(V)$, we know that

$$\mathrm{var}_{F_\alpha}(\Pi_{F_\alpha}[S_{\alpha_4}(\alpha) \,|S_{\alpha_3}(\alpha)]) \geq \mathrm{var}_{F_\alpha}\{\Pi_{F_\alpha}[S_\tau(\alpha) \,|\Lambda_3(0; \alpha)]\}.$$

However, by $S_{\alpha_3}(\alpha) \subset \Lambda_3(0; \alpha)$,

$$\mathrm{var}_{F_\alpha}(\Pi_{F_\alpha}[S_{\alpha_4}(\alpha) \mid S_{\alpha_3}(\alpha)]) \leq \mathrm{var}_{F_\alpha}\{\Pi_{F_\alpha}[S_\tau(\alpha) \mid \Lambda_3(0; \alpha)]\}.$$

Thus, $S_{\tau, \mathit{eff}}(\alpha) = S_{\tau, \mathit{eff}_{,par}}(\alpha)$.

Hence

$$\mathit{IF}_\tau(F_\alpha) = \mathit{IF}_{\tau, par}(\alpha) \text{ w.p. 1 for all } \alpha \in A(0). \tag{25}$$

Further, from its formula, when $\alpha_4 = 0$, $\mathit{IF}_{\tau, par}(\alpha)$ depends on α only through α_3 and on the data only through (Y, Z, V).

□

This lemma establishes premise (i) of Theorem 2. However, we cannot apply convexity to establish premise (ii) of Theorem 2 because the model $M(0)$ is not convex. This is because $Y \perp\!\!\!\perp Z \mid V$ is not preserved under convex combination. For example the convex combination of laws with densities $f_1(Y|V)f(Z)f_1(V)$ and $f_2(Y|V)f(Z)f_2(V)$ does not in general satisfy $Y \perp\!\!\!\perp Z \mid V$ unless $f_1(V) = f_2(V)$. Based on this observation, we consider the submodel M_v of M and $M_{v,sub,par}$ that assumes $f(v)$ equals a known density $f_0(v)$ and that the model $f(v;\alpha_2)$ satisfies $f(v;\alpha_2)|_{\alpha_2=0} = f_0(v)$. Note that this latter condition can always be arranged; since $f_0(v)$ is known, we simply choose any model $f(v;\alpha_2)$ that satisfies the condition. The model $M_v(0) = M_v \cap M(0)$ is convex since $f(z,v)$ is the same for all $F \in M_v(0)$. Specifically, under convex combinations, (a) $Y \perp\!\!\!\perp Z \mid V$ is preserved and (b) the constraints Eqs. (4) and (5) are also preserved, as the constraints are linear in $p(y,x|z,v)$. Furthermore, by inspecting the proof of Lemma 1, we see that $\mathit{IF}_{\tau}(F_{\alpha})$ and $\mathit{IF}_{\tau,par}(\alpha)$ under model $M_v(0)$ and $M_{v,sub,par}(0)$ and models $M(0)$ and $M_{sub,par}(0)$ are identical. Thus both premises of Theorem 2 hold for models M_v and $M_{v,sub,par}$. Hence, by Theorem 2, for $F \in M_v(0)$, $\alpha(F)$ solving $E_F[S(\alpha)] = 0$ subject to the constraint $\alpha_2 = 0$ required by model $M_{v,sub,par}$ satisfies $\widetilde{\tau}(\alpha(F)) \equiv \alpha_4(F) = 0$. However since $S_{\alpha_2}(\alpha)$ only depends on α through α_2 and $S_{\alpha\setminus\alpha_2}(\alpha)$ is not a function of α_2, we conclude that for $F \in M_v(0)$, $\alpha(F)$ solving $E_F[S(\alpha)] = 0$ without constraints also satisfies $\widetilde{\tau}(\alpha(F)) \equiv \alpha_4(F) = 0$. (The discussion in the last paragraph becomes unnecessary if, as is often assumed, we treat the distribution of V as fixed at its empirical, *i.e.*, we effectively condition on the observed values of V.)

But each $F \in M(0)$ is an element of a model $M_v(0)$; the model with $f_0(v)$ equal to the density of V under F. We conclude that for each $F \in M(0)$, $\alpha(F)$ solving $E_F[S(\alpha)] = 0$ satisfies $\widetilde{\tau}(\alpha(F)) \equiv \alpha_4(F) = 0$. This result holds therefore even when $f_0(v)$ is unknown and the chosen model $f(v;\alpha_2)$ is mis-specified. □

Rosenblum and Van Der Laan (2009) give an alternate proof of the fact that $S_{\tau,\mathit{eff}}(\alpha) = S_{\tau,\mathit{eff},\ \mathit{par}}(\alpha)$ that does not use Theorem 2.

To prove the analogous result for our ITT-null-robust parametrization for Mon_X + $\mathrm{Ex}_{\mathrm{AT}}$ + $\mathrm{Ex}_{\mathrm{NT}}$ it suffices to show that

$$\partial p_x^{z=i}(Y,V;\alpha\setminus\alpha_2)/\partial\alpha_4 = \partial p_x^{z=i}(Y,V;\alpha\setminus\alpha_2)/\partial\alpha_3 = 0, \quad i = 0,1 \text{ when } \alpha_4 = 0,$$

under our parametrization. Again our ITT-null-robust parametrization was carefully constructed to ensure that these derivatives were zero. The remainder of the proof is analogous to that for the robust parametrization of $\mathrm{Ex}_{\mathrm{AT}}$ + $\mathrm{Ex}_{\mathrm{NT}}$.

DISCUSSION

STEPHEN E. FIENBERG (*Carnegie Mellon University, USA*)

Overview. Given the by now extensive literature on non-compliance in randomized experiments, one might have thought that there was little new to be said, especially about the simplest such studies, when the idealized data come in the form of a $2 \times 2 \times 2$ contingency table. Further, given the commentary about the role of randomization for Bayesians that appears in at least part of the literature, one might presume that there is nothing that is necessarily Bayesian about the problem at all. Both of these presumptions are false.

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science in the Department of Statistics, the Machine Learning Department, Cylab, and i-Lab at Carnegie Mellon University, Pittsburgh PA 15213-3890, USA

Richardson, Evans, and Robins (henceforth RER) have written a stimulating paper, one worthy of careful study by Bayesians and non-Bayesians of all stripes and flavors, and not just by those interested in the non-compliance problem. RER have a new and remarkably clear message about this topic and this is the central idea in the paper, which I paraphrase:

> In many randomized experimental contexts the causal estimands of interest are not identified by the observed data. We need to re-parametrize partially identified models to separate wholly-identified and wholly-non-identified parameters. Our goal as statisticians is to focus on what is identifiable. As Bayesians, when we then look at such model structures, we need to recognize that what we will get out of our posterior distribution for the non-identifiable part will essentially be what we put into the prior, and unaffected by the experimental data!

Like prior authors addressing the non-compliance problem, RER use latent structure to resolve the identification of parameters and to make inference about causal effects and adopt many of the conventions for these found in the earlier literature. Some of the specifications for the latent structure have heuristic appeal but in many ways despite their continuing use are arbitrary. The impact of these choices, not surprisingly, plays an important role in what is identifiable and estimable.

Whether this is a message for Bayesians or for all statisticians and those who use statistical methods to analyze data from experiments subject to non-compliance is a topic to which I return below. In addition, in what follows I address the following pair of questions:

(i) How should we define causal effects?

(ii) Can we use algebraic geometry ideas to restructure the problem?

Latent structure and causal inference. The basic experimental structure RER describe takes the simple form of a $2 \times 2 \times 2$ contingency table with treatment variable Z, exposure variable X, and response variable y each of which is binary. Because they are dealing with a randomized experiment, the values of Z come in pre-specified form. These three random variables are linked via a graphical model. Since Y has both X and Z as parents there are four potential outcomes:

$$Y(x=0,z=0),\ Y(x=0,z=1),\ Y(x=1,z=0),\ Y(x=1,z=1).$$

The latent structure on Y, t_Y, involves classes of individuals or types. These don't have any kind of agreed-upon name in the literature but they get simplified with later assumptions. There is also a latent structure on exposure, X, which is unobservable involving four types: $\{NT, CO, DE, AT\}$.

When Y has only X as a parent, Z has no direct effect on Y, and there are just four Y-latent types, which by convention are typically called "Helped" $= (Y(x=0)=0, Y(x=1)=1)$, "Hurt," "Always Recover," and "Never Recover." Thus we have four Y-latent types and four X-latent types called "Never Takers," "Compliers," "Defiers," and "Always Takers" or 16 types for $\{t_X, t_Y\}$: $\{NT, CO, DE, AT\} \times \{HE, HU, AR, NR\}$, far more than the original counts in the 2^3 table about which we had hoped to make inferences. Something has to give way or the infrastructure will crash in a heap.

Finally, if no one in the placebo arm of the randomized trial can get the treatment (often not true), $Z = 0$ implies $X = 0$, and there are no "Defiers" or "Always Takers." Thus there are only two X-latent types and four Y-latent types: $\{NT, CO\} \times \{HE, HU, AR, NR.\}$ as in Figure 9. As RER note, this is the simple situation which arises for the Lipid data under Pearl's analysis (which presumes that Z does not have a direct effect on Y).

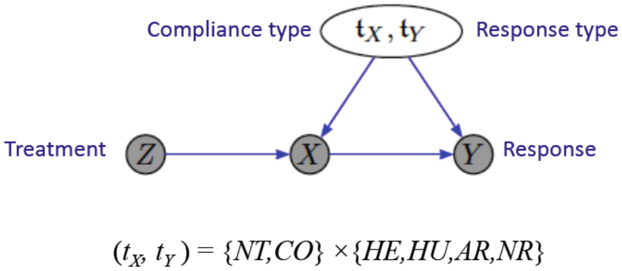


Figure 9: *Graphical representation of the Pearl IV model for Lipid data.*

At this point I'd like to offer a major caution. As appealing as the heuristic labels and categories for $\{t_X, t_Y\}$ are, we need to remember latent variables are simply that and thus unobservable. They are basically fictions which we introduce to help the modeling process. The true structure underlying compliance in a real randomized trial is obviously more complex and possibly not captured by the nice labels used in the compliance literature. We tend to forget this when we make graphical pictures such as in Figure 9. Therefore, we must proceed with caution when we move to the estimation of causal effects.

Once RER lay out this infrastructure for compliance, they spend most of the remainder of the paper working out plausible inequality restrictions that sharpen focus on $p(y, x|z)$. In particular they derive upper and lower bounds on causal relationships and "effects" under different assumptions about $\{t_X, t_Y\}$, specifically what they refer to as

(Mon_X) Monotonicity of compliance;
($\mathrm{Ex}_{\mathrm{NT}}$) Stochastic exclusion for NT under non-exposure;
($\mathrm{Ex}_{\mathrm{AT}}$) Stochastic exclusion for AT under exposure.

These bounds allow RER to talk about identification and estimability. Along the way, they include a variety of insights to the modeling process including a a way to represent the different models that are not inherently graphical using directed acyclic graphical ideas, albeit at the expense of some additional complexity. There is a richness of detail here worthy of study by those interested in the compliance problem.

In praise of Bayes? Should we be surprised to see a paper like this at *Valencia 9*? It does contain the naive Bayesian approaches to this problem, as I noted above, but to make the point that these just disguise the identification problem and can easily lead to nonsensical inferences. The only way to make inferences about essentially non-identifiable parts of the model is via strong information about them in the prior. RER make this point convincingly by example.

Most of bounding arguments and results could easily be presented from a frequentist perspective. In fact, Richardson and Robins (2010), in a companion paper to the present one, actually describe this problem and the basic ideas from a likelihood perspective. The real action in the present paper is about the specification of bounds on probabilities in the model and not so much on inference *per se.* There are implications of argument for Bayesian analyses, but I find them more cautionary than prescriptive and I would have liked to see the work culminate in a full-scale informative Bayesian analysis.

Yet there may be a deeper reason for Bayesians to take note of what RER have to say. After all, many of us have been long convinced by Rubin's (1978) argument about how randomization cuts the ties to covariates and allows for a direct assessment of the causal effect. RER are also exploiting randomization for the restrictions in the model and to garner identification, albeit in alimited way, but Bayesians are barely better off than frequentists when it comes to dealing with the part of the model that is not identifiable.

Alternative specifications for causal effects. RER follow prior authors and work with what is usually called the "Average Causal Effect":

$$ACE(X\ causes\ Y) = \log\left[\frac{E[Y_{x=1}]}{E[Y_{x=0}]}\right] = \log\left[\frac{\pi(help) + \pi(always\ recover)}{\pi(hurt) + \pi(always\ recover)}\right]$$

as well as "Intent to Treat Effect" (ITT) effects, principal stratification, and ACDE effects. It is as if these "causal" quantities were imbued with some objective status independent of the structure of the problem at hand. I think this is misleading.

Why not define causal effect based on $\log(E[Y_{x=1}]/E[Y_{x=0}])$? If we used this alternative specification for causal effect, we would be working with adjusted odds ratios or ratios of odds ratios for the basic 2×2 table, and then the definition of the causal effect would be rooted in a logit-like statistical model instead of the linear model implicit in the definition of ACE, ITT, and ACDE. Sfer (2005) develops this kind of argument for a simple randomized experiment involving a binary treatment variable and a binary outcome. My guess is that much of the thinking in the present paper would carry over to this model-based representation.

In the companion paper by Richardson and Robins, the authors derive much of inequality results using geometry arguments. In many ways, I missed this elegant geometric representation in the present paper. I'd be interested in exploring how the restrictions derived in this pair of papers effect the characterization of $p(x, y, z)$ and thus $p(x, y|z)$, through marginals, conditionals, and ratios of odds ratios for the alternative definition of causal effect based on odds ratios and log-linear parameters. Then we might be able to exploit algebraic geometric arguments such as those described in Slavković and Fienberg (2010). This was a problem I tried to address, with singular lack of success, when I first saw the bounds in Balke and Pearl (1997) thirteen years ago. Perhaps it is time to return to it with RER's work as a guide.

PAUL GUSTAFSON (*University of British Columbia, Canada*)

I congratulate Richardson, Evans and Robins (hereafter RER) on a very interesting paper. I completely agree that given a partially identified problem, identifying a transparent parameterization is key to understanding the efficacy of Bayesian inference. I also think that the paper breaks fruitful new ground in exploring Bayesian inference for the instrumental variables model based on potential outcomes. Given the ease with which a Bayesian solution integrates uncertainty due to finite sampling and uncertainty due to a lack of identification, consideration of Bayesian inference for potential outcome models seems important in general.

One curious point about partially identified models is that in the asymptotic limit the Bayesian has more to convey than the frequentist. That is, both will agree on the set of possible values for the target parameter, but the Bayesian will additionally weight the plausibility of different values in this set with respect to one another, *i.e.*, the limit of the posterior distribution will have some shape. It seems relevant to ask about the utility of this shape, and the extent to which it is driven by the data versus the prior.

As a simple illustration, consider a slight extension of RER's motivating example in Section 3. Still making the "randomized trial" assumption $Y_{x0} = Y_{x1}$, consider the three compliance types $\{NT, AT, CO\}$, *i.e.*, always-takers, but not defiers, have been added to the mix. Following the RER notation, the situation can be understood scientifically with reference to parameterization

$$(\pi_{NT}, \pi_{AT}, \gamma_{CO}^{0\cdot}, \gamma_{CO}^{1\cdot}, \gamma_{NT}^{0\cdot}, \gamma_{NT}^{1\cdot}, \gamma_{AT}^{0\cdot}, \gamma_{AT}^{1\cdot}),$$

where implicitly $\pi_{CO} = 1 - \pi_{NT} - \pi_{AT}$. These scientifically interpretable parameters can simply be cleaved into a wholly identified component

$$\phi = (\pi_{NT}, \pi_{AT}, \gamma_{CO}^{0\cdot}, \gamma_{CO}^{1\cdot}, \gamma_{NT}^{0\cdot}, \gamma_{AT}^{1\cdot}),$$

and a component $\psi = (\gamma_{NT}^{1\cdot}, \gamma_{AT}^{0\cdot})$ which is not involved in the likelihood function.

Now, say the target of inference is the average causal effect,

$$ACE \quad = \quad (1 - \pi_{NT} - \pi_{AT})(\gamma_{CO}^{1\cdot} - \gamma_{CO}^{0\cdot}) + \pi_{NT}(\gamma_{NT}^{1\cdot} - \gamma_{NT}^{0\cdot}) + \pi_{AT}(\gamma_{AT}^{1\cdot} - \gamma_{AT}^{0\cdot}).$$

Thus, regardless of whether one pursues a frequentist or Bayesian analysis, in the asymptotic limit one learns the range of possible values for the target is $a \pm b$, where

$$a \quad = \quad (1 - \pi_{NT} - \pi_{AT})(\gamma_{CO}^{1\cdot} - \gamma_{CO}^{0\cdot}) + \pi_{NT}(1/2 - \gamma_{NT}^{0\cdot}) + \pi_{AT}(\gamma_{AT}^{1\cdot} - 1/2),$$

and $b = (\pi_{NT} + \pi_{AT})/2$.

From a frequentist viewpoint, (a, b) are all that can be learned about the ACE from an infinite-sized dataset. The situation is different, however, for a Bayesian. The large-sample limit of the posterior distribution must have $a \pm b$ as its support, but additionally the shape of the limiting distribution may depend on the identified parameters.

For instance, say a uniform prior is applied in the scientific parameterization (more formally a Dirichlet$(1, 1, 1)$ prior for $(\pi_{NT}, \pi_{AT}, 1 - \pi_{NT} - \pi_{AT})$, and Unif$(0, 1)$ priors for each of γ's, with independence throughout). Then it follows directly that the large-sample limit of the posterior distribution for the ACE has a stochastic representation as $a + \pi_{NT}(U_1 - 1/2) + \pi_{AT}(U_2 - 1/2)$, where U_1, U_2 are *iid* Unif$(0, 1)$. It then follows that the limiting distribution is symmetric on $a \pm b$, with a trapezoid-shaped density. Particularly, the top edge of the trapezoid extends along $a \pm (\max\{\pi_{NT}, \pi_{AT}\} - \min\{\pi_{NT}, \pi_{AT}\})/2$. Extreme cases are a uniform limiting density (when $0 = \min\{\pi_{NT}, \pi_{AT}\} < \max\{\pi_{NT}, \pi_{AT}\}$), and a triangular limiting density (when $\pi_{NT} = \pi_{AT} > 0$). Thus, the peakedness of the limiting distribution depends on identified parameters, showing that the shape is not merely a pre-ordained consequence of the shape of the prior. As a practical implication, the width of a central credible interval for the target, relative to the width of the set of plausible values, varies according to the underlying true parameter values.

In fact, in the present problem the influence of the identified parameters on the shape of the posterior on the target is fairly mild. Particularly, the limiting posterior distribution is symmetric on $a \pm b$, no matter what. However, Gustafson (2010) gives examples of partially identified models where the shape depends more strongly on the identified parameters. This is more prone to occur when the identified parameter vector ϕ is a complicated function of the original scientific parameters, rather than simply a subset of them as in the example above.

FABRIZIA MEALLI (*Università di Firenze, Italia*) and
FAN LI (*Duke University, USA*)

We appreciated the invited paper by Richardson, Evans and Robins (henceforth RER), as the only one in Valencia 9 dealing with causal inference, an important branch of statistical inference for which Bayesian analysis is particularly and naturally suited (Rubin, 1978). However, in our view, the paper misses important recent advances made in causal inference, and specifically, in Bayesian analysis of broken randomized experiments. We would like to stress that a framework, namely Principal Stratification (PS; Frangakis and Rubin, 2002), exists that allows one to transparently specify causal models, to separate structural behavioral assumptions from model assumptions and priors on parameters, and to conduct model-based Bayesian inference in a principled fashion; the area could certainly benefit from cross-fertilization.

PS has been successfully applied to a wide range of more general and complicated settings, where the applicability of RER's approach is not completely clear. The aim of our discussion is three-fold: (i) to provide a brief account of the existing literature on the subject of Bayesian causal inference with intermediate variables; (ii) to elucidate how Bayesian inference is conducted under the PS framework; and (iii) to discuss some inferential and practical restrictions embedded in RER.

The all-or-none noncompliance setting analyzed by RER is an example of causal analysis with intermediate variables, that is, post-treatment variables potentially affected by treatment and also affecting the response. Much of the notation and terminology in RER stems from the series of papers by Imbens, Rubin and coauthors in the 1990s (*e.g.*, Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Imbens and Rubin, 1997), which provided the terminology of NT, AT, CO, and DE. Further, Frangakis and Rubin (2002) proposed the general PS framework for adjusting for intermediate variables, based on stratifying units by their joint potential intermediate outcomes. Since then, advances have been achieved in causal inference under PS, both from frequentist and Bayesian perspectives (the literature listed below is limited to the Bayesian one), dealing with settings of binary, categorical, continuous, censored outcomes with and without covariates (*e.g.*, Hirano, 2000; Zhang *et al.*, 2008); noncompliance coupled with missing data or/and censored data (*e.g.*, Barnard et al.; 2003; Mattei and Mealli, 2007); longitudinal treatments and intermediate variables (Frangakis *et al.*, 2004); clustered treatments (*e.g.*, Frangakis *et al.*, 2002); surrogate endpoints (Li *et al.*, 2010); continuous intermediate variables, including partial compliance (*e.g.*, Jin and Rubin, 2008; Schwartz *et al.*, 2010); just to name a few.

While RER provide some insights on the information which can be drawn from randomized experiments with noncompliance, their statement about “standard Bayesian prior to posterior analysis” of “weakly identified” models may suggest to those who are unfamiliar with the causal inference literature that the current state-of-the-art Bayesian analysis of such models is not done properly, or done in a rather “automatic” fashion, without posing attention on the nature of the different causal estimands, and on the information provided by the data on them. However, Imbens and Rubin (1997) already provided a complete recipe for model-based Bayesian inference and investigated the behavior of weakly identified models for the case of all-or-none compliance. Their approach can be easily generalized to conduct analysis with other intermediate variables, as briefly described below.

Let Z_i be the binary variable indicating the treatment assignment of unit i. Under SUTVA, the potential outcomes are a function of Z_i rather than the entire vector $\boldsymbol{Z}$. Let $Y_i(z)$ and $X_i(z)$ be the potential primary and intermediate outcomes if unit i is assigned to treatment z for $z = 0, 1$. In Bayesian inference, the observable quantities for a sample of N units, $(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1), \boldsymbol{Z}, \boldsymbol{V})$, are considered as observed and unobserved realizations of random variables, with joint distribution $\Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1), \boldsymbol{Z}, \boldsymbol{V})$ which may be written as

$$\begin{aligned}&\Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1)|\boldsymbol{Z}, \boldsymbol{V}) \Pr(\boldsymbol{Z}|\boldsymbol{V}) \Pr(\boldsymbol{V})\\ &\qquad\qquad = \Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1)|\boldsymbol{V}) \Pr(\boldsymbol{Z}|\boldsymbol{V}) \Pr(\boldsymbol{V}),\end{aligned}$$

where the equality follows from randomization, which allows one to separate the joint distribution of the potential outcomes from the treatment assignment mechanism. Analysis is usually conditional on the observed distribution of covariates, thus $\Pr(\boldsymbol{V})$ is not modelled. The joint distribution of the potential outcomes, $\Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1)|\boldsymbol{V})$, can be rewritten as

$$\int \prod_i \Pr(Y_i(0), Y_i(1)|X_i(0), X_i(1), V_i, \boldsymbol{\theta}) \Pr(X_i(0), X_i(1)|V_i, \boldsymbol{\theta}) \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \qquad (26)$$

for the global parameter $\boldsymbol{\theta}$ with prior distribution $\pi(\boldsymbol{\theta})$. The quantity $(X_i(0), X_i(1))$ is called a principal stratum S_i and the cross-classification of units into the latent classes of S_i is called PS. Clearly the classification of units into NT, AT, CO, DE is a special case of PS. The key insight is that S_i is invariant under different treatment assignments, thus the comparisons of $\{Y_i(1) : S_i = (x_0, x_1), V_i = v\}$ and $\{Y_i(0) : S_i = (x_0, x_1), V_i = v\}$ are well-defined causal effects (called principal causal effects—PCEs). Factorization (26) suggests that model-based PS inference usually involves two sets of models: One for the distribution of potential outcomes $Y(0), Y(1)$ conditional on the principal strata and covariates and one for the distribution of principal strata conditional on the covariates. The definition of principle strata does not involve response Y, unlike the approach in RER. We find it hard to envision how the approach in RER can be extended to, for example, the most common case of continuous Y.

Another critical feature in the models adopted in the PS literature is that one specifies models directly on potential outcomes instead of on observed quantities. This, we think, is a more transparent way of modelling, and also, by doing so we can “directly” check which restrictions are supported by the data.

Since causal effects are defined as (summaries of) comparisons between the potential outcomes of the same individuals, in our opinion directly modelling potential outcomes enables analysts to transparently conduct causal inference. To conduct Bayesian inference in PS, the complete-data likelihood of all units can be written

$$\begin{aligned}&\Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}(0), \boldsymbol{X}(1)|\boldsymbol{V}; \boldsymbol{\theta})\\ &\quad= \prod\nolimits_{i=1}^{N} \Pr\left(Y_i(0), Y_i(1)|S_i, V_i; \boldsymbol{\beta}^Y\right) \Pr\left(S_i|V_i; \boldsymbol{\beta}^S\right),\end{aligned}$$

where $\boldsymbol{\theta}$ includes parameters $\boldsymbol{\beta}^Y$'s and $\boldsymbol{\beta}^S$. The Bayesian model is completed by specifying prior distributions for each set of parameters. Note that separating the parameters into an identifiable set and a non/weakly identifiable set, as RER do, may lead to prior independence between parameters that are substantively thought to be closely related.

Define

$$\begin{aligned}Y_i^{obs} &= Y_i(Z_i),\\ Y_i^{mis} &= Y_i(1 - Z_i),\\ X_i^{obs} &= X_i(Z_i),\\ X_i^{mis} &= X_i(1 - Z_i).\end{aligned}$$

All PCEs are functions of the model parameters $\boldsymbol{\theta}$ and observed quantities, so full Bayesian inference for PCEs is based on the posterior distribution of the parameters conditional on the observed data, which can be written as,

$$\begin{aligned}&\Pr(\boldsymbol{\theta}|\boldsymbol{Y}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{Z}, \boldsymbol{V})\\ &\quad\propto \Pr(\boldsymbol{\theta}) \int\int \prod_i \Pr(Y_i(0), Y_i(1), X_i(0), X_i(1)|V_i, \theta) dY_i^{mis} dD_i^{mis}.\end{aligned}$$

However, direct inference from the above distribution is in general not available due to the integrals over D_i^{mis} and Y_i^{mis}. But both $\Pr(\theta|\boldsymbol{Y}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{X}^{mis}, \boldsymbol{V})$ and $\Pr(\boldsymbol{X}^{mis}|\boldsymbol{Y}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{V}, \boldsymbol{\theta})$ are generally tractable, so the joint posterior distribution, $\Pr(\boldsymbol{\theta}, \boldsymbol{X}^{mis}|\boldsymbol{Y}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{Z}, \boldsymbol{V})$, can be obtained using a data augmentation approach for $\boldsymbol{X}^{mis}$. Inference for the joint posterior distribution then provides inference for the marginal posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{Y}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{Z}, \boldsymbol{V})$.

Note that, by adopting PS, it is rather obvious that some parameters are "wholly" nonidentified. They are usually those which depend on potential outcomes that are never observed in a particular experiment for certain types of subjects (principal strata). For example Y_{11} is never observed for units (NT and DE) with $X_i(Z_i = 1) = 0$; it is thus an "a priori" counterfactual. As a consequence the effect of treatment receipt for NT does not appear in the likelihood, and no prior is put on it. Indeed, even after a Bayesian analysis, bounds can be derived on these quantities, *e.g.*, by letting the a priori counterfactual outcomes range from their smallest to their largest possible values (see, *e.g.*, Imbens and Rubin, 1997, page 319). More precise inference can be obtained, *e.g.*, by "extrapolation", assuming the effect found for CO is the same as the effect that would have been observed for NT, had they been forced to take the treatment. Also, because the effect on CO is well identified only under some restrictions, the Bayesian PS approach allows one to directly and transparently check what restrictions are supported by the data, by relaxing those restrictions and checking how much posterior support they receive.

REPLY TO THE DISCUSSION

We thank the discussants for their thoughtful comments and criticisms. We comment on each contribution in turn. (Readers should note that Section 6 and the appendix concerning mis-specification were added subsequently, hence were not seen by the discussants.)

Different causal effect measures. Fienberg notes that all of the causal contrasts in our paper are defined on the linear scale, and asks to what extent the methods can be applied to the causal relative risk or causal odds ratio. Richardson and Robins (2010) characterize the set of possible values for

$$(\pi_{\mathrm{NT}}, \pi_{\mathrm{AT}}, \pi_{\mathrm{DE}}, \pi_{\mathrm{CO}}, \gamma^{0\cdot}_{\mathrm{CO}}, \gamma^{1\cdot}_{\mathrm{CO}}, \gamma^{0\cdot}_{\mathrm{NT}}, \gamma^{1\cdot}_{\mathrm{NT}}, \gamma^{0\cdot}_{\mathrm{AT}}, \gamma^{1\cdot}_{\mathrm{AT}}, \gamma^{0\cdot}_{\mathrm{DE}}, \gamma^{1\cdot}_{\mathrm{DE}}),$$

compatible with a given observed population distribution $p(y, x \mid z)$ under the model $\mathrm{EX}_{\mathrm{AT}} + \mathrm{EX}_{\mathrm{NT}}$. Given this description it is straightforward to compute bounds on any causal contrast for this model regardless of the chosen scale. The methods of analysis may be extended fairly easily to any of the other models we consider, as we did to compute the bounds on $\mathrm{ITT}_{\mathrm{CO}}$ given in Figure 7.

More generally, we agree with Fienberg that algebraic geometry and, in particular, the theory of convex polytopes have an important role to play in understanding identifiability in similar potential outcome models in which variables have more than two states.

The role of re-parametrization. We thank Gustafson for pointing out that the use of a transparent re-parametrization in no way precludes the specification of a prior on the non-identified parameters. From our perspective, re-parametrization is purely a mathematical technique for clarifying the relationship between the data and the posterior by separating the wholly identified from the wholly unidentified. More formally, let θ be the vector of parameters in the original (*e.g.*, "Principal" Stratum) formulation, and let $(\psi, \aleph)$ indicate the transparent re-parametrization into identified (ψ) and unidentified ($\aleph$) components, via some diffeomorphism $g(\cdot)$, so $\theta = g(\psi, \aleph)$. A prior $p(\theta)$ induces a prior $p(\psi, \aleph) = p(\psi)p(\aleph \mid \psi)$. Then we have:

$$p(\psi, \aleph \mid y) = p(\aleph \mid \psi)p(\psi \mid y), \tag{27}$$

where we have used the fact that $\aleph$ does not occur in the likelihood. Forward sampling may be used to obtain samples from the posterior $p(\theta \mid y)$ by first sampling $\psi^{(i)}$ from $p(\psi \mid y)$ and then sampling $\aleph^{(i)}$ from $p(\aleph \mid \psi^{(i)})$. The corresponding value of $\theta^{(i)} = g(\psi^{(i)}, \aleph^{(i)})$.

In the illustrative analyses we present in the paper (that make use of re-parametrizations) we avoided placing prior distributions on parameters that were not identified, instead opting to compute posterior distributions on bounds. This was primarily because we thought that in many circumstances useful subjective information relating to these specific unidentified quantities may be hard to come by. We make a few further points in this regard below.

Transparent re-parametrization is, in principle, compatible with any Bayesian analysis of a partially identified model. We say "in principle" because it may require some technical work to be able to find such a re-parametrization: this is one of the main contributions of our paper for the unidentified models we consider.

However, it is therefore incorrect to suggest, as Mealli and Li do, that an analyst must choose between a "PS approach" and our re-parametrization. Like co-ordinate systems, an analyst is free to use more than one parametrization within a single analysis: if background knowledge is more amenable to formulation via reference to compliance types then the prior may be formulated in these terms. However, we, like Gustafson, Greenland and Leamer before us, believe that in order to assess the extent to which beliefs regarding unidentifiable quantities influence the posterior distribution it is often necessary to use a transparent re-parametrization.

Relation to existing methods. Mealli and Li state that Imbens and Rubin (1997) provide a complete recipe for model-based Bayesian inference of "broken" randomized experiments with non-compliance. We do not agree. We believe that the method of Imbens and Rubin is incomplete in its treatment of two central issues: (i) the sensitivity of the posterior for a partially identified quantity to the prior; (ii) bias under model mis-specification in randomized experiments. We now consider each in turn.

Prior sensitivity.

In order to eliminate extraneous issues, we assume baseline covariates are either absent, or take only a few values.

Inference for partially identified quantities vs. inference for bounds. An important theme in our paper is that when faced with a partially identified parameter, it is advisable to proceed by computing the bounds on this quantity implied by the population distribution for the observables, and then to perform inference for these bounds. Such bounds, being functionals of the observed distribution, are identified. Though we did not stress this point in the paper, such an analysis need not preclude, and indeed may complement, a standard Bayesian analysis for the quantity of interest. Thus, contrary to Mealli and Li, it is a false dichotomy to suggest that an analyst must choose one or the other.

Testing exclusion restrictions. Hirano *et al.* highlight the ability to relax individual exclusion restrictions as one of the strengths of their approach. In their remarks Mealli and Li write:

> ...the Bayesian PS approach allows one to directly and transparently check what restrictions are supported by the data, by relaxing those restrictions and check[ing] how much posterior support they receive.

In our opinion, this remark indicates the danger of Bayesian analyses that fail to distinguish what is from what is not identifiable.

To see this point more clearly, consider the restriction Ex_{AT}, which is one of the assumptions necessary to identify ITT_{CO}. Note that the assumption Ex_{AT} is equivalent to $\text{ITT}_{\text{AT}} = 0$. The quote of Mealli and Li above suggests that, had a 99.9% credible interval for ITT_{AT} excluded zero, Mealli and Li would regard this as overwhelming evidence that Ex_{AT} is false, even if they (following Hirano *et al.*) had used "off-the-shelf" priors. But such an inference would be erroneous. It is possible that the (identifiable) population lower and upper bounds, denoted by $l\text{ITT}_{\text{AT}}$ and $u\text{ITT}_{\text{AT}}$, for ITT_{AT} straddle zero, yet owing to the specific prior used, the posterior credible interval for ITT_{AT} may exclude zero; compare to the right panel of Figure 3. In contrast, an analysis based on credible intervals for bounds will (asymptotically) not make such a mistake; see Table 9, with Ex_{AT} true.

Hirano *et al.* concluded, based primarily on subject matter considerations, that there was reason to doubt the exclusion restriction for Always Takers in the McDonald *et al.* data. (However, this decision was not a consequence of their likelihood or "off-the-shelf" prior.) Our Bayesian inference for bounds provides some empirical support for this doubt; see Column (13) in Table 7.

Table 9: *Contrasting large-sample inferences for* ITT$_{\text{AT}}$ *vs. large-sample inferences for upper and lower bounds on* ITT$_{\text{AT}}$*.* PD *indicates that this is prior dependent (even asymptotically);* uITT$_{\text{AT}}$ *and* lITT$_{\text{AT}}$ *are the (identifiable) upper and lower bounds. Mon*$_X$ *is assumed to hold.* LSCI *is a credible interval in large samples.*

	True State		
	Ex$_{\text{AT}}$ true	Ex$_{\text{AT}}$ false	
Result of Posterior Analysis	(13) true	(13) true	(13) false
$0 \in$ LSCI[ITT$_{\text{AT}}$]	PD	PD	False
$0 \notin$ LSCI[ITT$_{\text{AT}}$]	PD	PD	True
LSCI[uITT$_{\text{AT}}$] $\cap [0,\infty) \neq \emptyset$ **and** LSCI[lITT$_{\text{AT}}$] $\cap (-\infty,0] \neq \emptyset$	True	True	False
LSCI[uITT$_{\text{AT}}$] $\cap [0,\infty) = \emptyset$ **or** LSCI[lITT$_{\text{AT}}$] $\cap (-\infty,0] = \emptyset$	False	False	True

Table 10: *Parameter values for simulation scenarios: all linear models* (L)*; quadratic compliance model* (QC)*; quadratic qesponse models* (QR)*.*

	α_0	α_1	α_2	β_0^c	β_1^c	β_2^c	δ_0^c	δ_1^c	β_0^n	β_1^n	β_2^n
L	2	−0.3	**0**	−2	1	**0**	**0**	**0**	5	−2	**0**
QC	3.5	−2.5	0.5	−2	1	**0**	**0**	**0**	5	−2	**0**
QR	2	−0.3	**0**	4.6	−4.8	1	**0**	**0**	3	−5	1.2

Examining prior sensitivity. The question of the extent to which prior specification influences the posterior may arise in any Bayesian analysis. However, we believe that *ad hoc* approaches, which may be appropriate in identified contexts, such as finding "equivalent sample sizes" or comparing the prior and posterior standard deviations for quantities of interest (see Hirano *et al.*, 2000, p. 78), may be highly misleading in the context of partially identified parameters and are not logically justified.

As an example, suppose the population bounds on a partially identified parameter of interest were $(-2, 2)$. If a Bayesian analyst specified a diffuse but proper prior with a 99% credible interval of $(-50, 50)$ then in large samples, the posterior standard deviation will be at most 1/20 of the prior standard deviation (under some assumptions on the shapes of the prior and posterior). Nevertheless, owing to the sensitivity of the posterior to even a diffuse prior, the Bayesian's posterior 99% credible interval could be, $(-1.8, -0.3)$ even though the true value of the partially identified parameter was 0.5; see Figure 3.

"Principal" strata vs. strata defined by baseline covariates. One of the primary motivations for the principal stratum (PS) approach is that, if the strata are based on

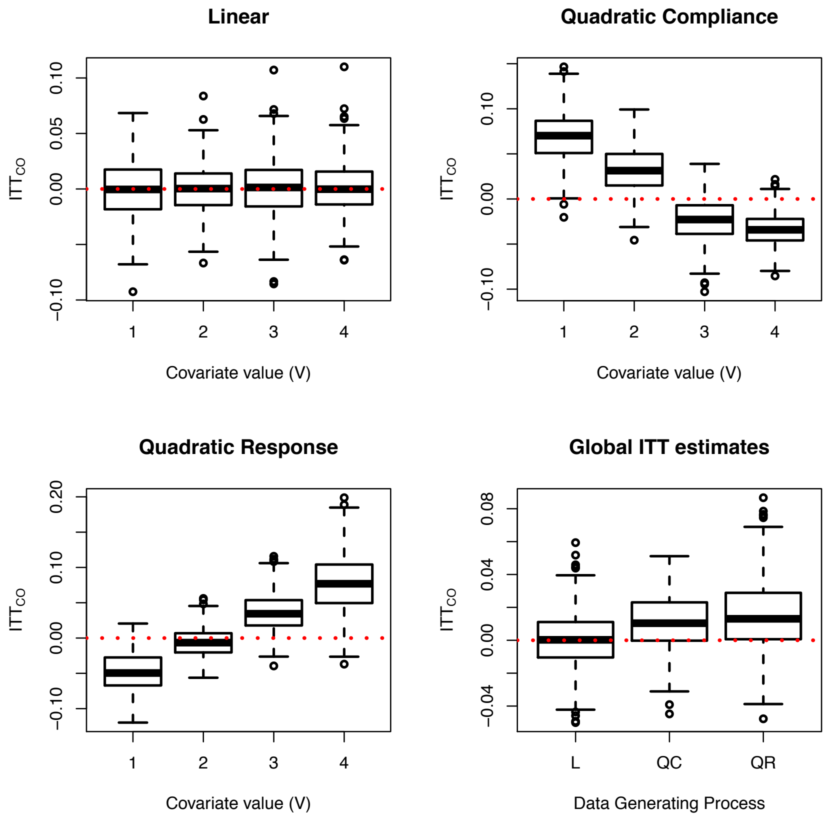


Figure 10: *Boxplots showing the sampling distribution of the MLE for* $\mathrm{ITT}_{\mathrm{CO}}(v)$*: linear (top left); quadratic compliance (top right); quadratic response (bottom left). Sampling distributions of the MLE for the global* $\mathrm{ITT}_{\mathrm{CO}}$ *(bottom right). The true values of all* $\mathrm{ITT}_{\mathrm{CO}}$ *parameters are zero; see dotted line.*

well-defined potential outcomes for the intermediate, contrasts between the treated and untreated within such strata admit a causal interpretation. In this regard, principal strata are analogous to a set of baseline covariates sufficient to control confounding. However, in another important respect, principal strata are very different from baseline covariates in that, in general, we are never able to directly observe such memberships. Consequently, prior information regarding differences in response between compliance types are likely to be scarce and unreliable. This is a major concern in light of the extreme sensitivity of the posterior distribution for weakly identified parameters to the choice of prior.

Model mis-specification in randomized experiments

We now consider the consequences of model mis-specification for the method of Hirano *et al.* in a randomized experiment with continuous baseline covariates V. Inclusion of baseline covariates in the analysis is useful because qualitative treatment-covariate interactions can be detected and, as noted by Hirano *et al.*, efficiency may be increased.

We now consider the setting of a double blind (DB) placebo-controlled RCT in which treatment is without side-effects and is not available to patients in the control arm. In this setting Defiers and Always Takers are not present. Furthermore, the exclusion restrictions for Never Takers ($\mathrm{EX}_{\mathrm{NT}}$) and for Compliers ($\gamma^{00}_{\mathrm{CO}} = \gamma^{01}_{\mathrm{CO}}$, $\gamma^{10}_{\mathrm{CO}} = \gamma^{11}_{\mathrm{CO}}$) can be assumed to hold within levels of V. Then the conditional intent to treat effect $\mathrm{ITT}_{\mathrm{CO}}(V)$ in the Compliers is identified by

$$\frac{\mathrm{ITT}_Y(V)}{\mathrm{ITT}_X(V)} \equiv \frac{E\,[Y \mid Z=1,V] - E\,[Y \mid Z=0,V]}{E\,[X \mid Z=1,V] - E\,[X \mid Z=0,V]}$$

and equals the conditional Complier Average Causal Effect $\mathrm{ACE}_{\mathrm{CO}}\,(X \to Y \mid V)$ of X on Y. The unconditional ITT effect, $\mathrm{ITT}_{\mathrm{CO}} = E\,[\mathrm{ITT}_{\mathrm{CO}}(V)]$ and unconditional Complier Average Causal Effect are also identified.

In Section 6, we described why it is critical to analyze randomized trials with a method that, under the null hypothesis that $\mathrm{ITT}_Y(V) = 0$, guarantees that the posterior distribution and MLE of $\mathrm{ITT}_{\mathrm{CO}}(V)$ [and thus of $\mathrm{ITT}_Y(V)$] concentrate on the zero function, even under model mis-specification. The following simulation study demonstrates that the method of Hirano *et al.* does not offer such a guarantee, even when V is discrete.

We simulated the data under the model:

$$\begin{aligned}
p(\mathfrak{t}_X = \mathrm{CO} \mid v, z) &= \mathrm{expit}(\alpha_0 + \alpha_1 v + \alpha_2 v^2),\\
p(Y = 1 \mid \mathfrak{t}_X = \mathrm{CO}, v, z) &= \mathrm{expit}(\beta^c_0 + \beta^c_1 v + \beta^c_2 v^2 + z(\delta^c_0 + \delta^c_1 v)),\\
p(Y = 1 \mid \mathfrak{t}_X = \mathrm{NT}, v, z) &= \mathrm{expit}(\beta^n_0 + \beta^n_1 v + \beta^n_2 v^2),
\end{aligned}$$

with $\delta^c_0 = \delta^c_1 = 0$ hence $\mathrm{ITT}_Y(V) = 0$. The baseline covariate V is ordinal, distributed uniformly with sample space $\{1, 2, 3, 4\}$. Always Takers and Defiers were excluded *a priori*. We considered data simulated under three different parameter settings as shown in Table 10; in the first (L), there are no quadratic terms; in the second (QC), there is a quadratic term in the logistic regression model for the proportion of Compliers vs. Never Takers; in the third (QR), there is a quadratic term present in the logistic regression models for $E[Y \mid X = 0, \mathfrak{t}_X]$ for $\mathfrak{t}_X \in \{\mathrm{CO}, \mathrm{NT}\}$. For each scenario we simulated 500 datasets of size 5,000. We used the linear logistic model of Hirano *et al.* without Always Takers to analyze the data; we purposely omitted the quadratic terms from the models fitted. Since we are interested primarily in large sample performance we used the MLE and standard asymptotic 95% Wald confidence intervals as convenient approximations to the posterior mode and 95% credible intervals. Table 11 gives the sampling distribution of the MLEs for δ^c_0 and δ^c_1, together with the actual coverage rate for nominal 95% and 90% asymptotic confidence intervals. Figure 10 shows sampling distributions of the MLEs for $\mathrm{ITT}_{\mathrm{CO}}(V)$ and $\mathrm{ITT}_{\mathrm{CO}}$ under each of these scenarios. As can be seen, misspecification of either the model for compliance types or for $E[Y \mid X = 0, \mathfrak{t}_X]$ leads to spurious inferences regarding the ITT effects even under the ITT null.

As discussed in our paper, the model (the last in Table 6), that we had proposed in Section 5 to analyze such a trial also failed to satisfy the wished-for guarantee; see the simulation study in §6. As noted earlier, Mealli *et al.* wrote their discussion based on an earlier version of the paper that did not include Section 6; hence they had no opportunity to express their thoughts on this issue.

Table 11: *Simulation results: distribution of MLE for δ_0^c and δ_1^c fitting potential outcome models omitting quadratic terms. Coverage shows the actual coverage corresponding to asymptotic confidence intervals based on the observed information. Results based on* 500 *simulations; sample size was* 5, 000*. The ITT null holds so* $\delta_0^c = \delta_1^c = 0$.

	δ_0^c MLE		Coverage		δ_1^c MLE		Coverage	
	Mean	Std. Err	90%	95%	Mean	Std. Err	90%	95%
L	−0.001	(0.010)	0.91	0.94	0.000	(0.005)	0.91	0.94
QC	0.331	(0.005)	0.20	0.31	−0.257	(0.005)	0.22	0.35
QR	−0.362	(0.008)	0.31	0.43	0.169	(0.003)	0.25	0.34

As noted in the paper, we specifically developed the parametrization and parametric model described in the last subsection of §6 to provide robustness to model mis-specification under the ITT null. Note, however, that this parametric model allows for the possibility of Always Takers. A robust model that assumes the absence of Always Takers is obtained by simply setting the functions $\kappa_i(v; \alpha_1)$, $i \in \{0, 1\}$, to zero in the aforementioned model.

Our simulation study demonstrated, by example, the non-robustness of the Hirano *et al.* approach in the simple setting of a single V with only four levels and a sample size of 5,000; see Table 11 and Figure 10. As such, it is likely that various goodness-of-fit statistics would reject the linear analysis model with high power. However, when V is a vector with continuous components, it is more difficult to specify correct or nearly correct parametric models and the power of goodness-of-fit statistics to reject even a quite mis-specified model is poor. Thus we suspect that in high-dimensional settings the use of non-robust parametric models will typically result in markedly incorrect inference under the conditional ITT null.

Summary

We certainly do not claim that all model-based analyses of partially-identified quantities are equally misleading. Indeed, we found the Hirano *et al.* Bayesian analysis to be interesting, thoughtful, and restrained in its conclusions.

In contrast, Zhang *et al.* (2009) analyze partially-identified direct effects in a job-training program. They entirely eschew a Bayesian approach, preferring instead to (i) specify a parsimonious parametric model whose functional form serves to point-identify the direct effects, and (ii) estimate these effects by maximum likelihood (without associated standard errors). They summarize their inferences with bold pronouncements such as: “there is a group of individuals, about 8%, for whom assignment to training is harmful in terms of employment,” without any measure of uncertainty.

We can summarize our concerns by echoing David Freedman’s invocation of Will Rogers, who famously said: “It’s not what you don’t know that hurts. It’s what you know that ain’t so. . . ,” the model-based approach advocated by Mealli *et al.* is likely to “increase the stock of things we know that you know for sure that just ain’t so.”[1]

Finally, our paper and those referenced by Mealli and Li were concerned with making inferential statements about causal effects of scientific interest and not with

[1] Sander Greenland notes that this quote is originally due to Mark Twain.

decision-making under uncertainty. When a decision must be made and the optimal choice depends upon an unknown partially-identified effect parameter, we, like all Bayesians, would use our personal posterior distribution for the parameter of interest. In such a situation we would use our proper subjective prior; we would not use either a default or "off-the-shelf" prior.

ADDITIONAL REFERENCES IN THE DISCUSSION

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91**, 444–455.

Barnard, J., Frangakis, C. F., J., Hill, J. L., Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *J. Amer. Statist. Assoc.* **98**, 299–323 (with discussion).

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002). Clustered encouragement design with individual noncompliance: Bayesian inference and application to advance directive forms. *Biostatistics* **3**, 147–164.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *J. Amer. Statist. Assoc.* **99**, 239–249.

Gustafson, P. (2010). Bayesian inference for partially identified models. *Internat. J. Biostatistics* **6**, Art 17.

Imbens, G. W., Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–476.

Jin, H., Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103**, 101–111.

Li, Y., Taylor, J. M., Elliott M. R. (2009). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**, 523–531.

Mattei, A., Mealli, F. (2007). Application of the prinicipal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63**, 437–446.

Robins, J., Rotnitzky, A., and Vansteelandt, S. (2007) Discussion of *Principal stratification designs to estimate input data missing due to death* by Frangakis, C.E., Rubin, D.B., An, M., MacKenzie, E. *Biometrics* **63**, 650–653.

Rubin, D. B. (1978). Bayesian inference for causal effects. *Ann. Statist.* **6**, 34–58.

Schwartz, S.L., Li, F., and Mealli, F. (2010). A Bayesian semiparametric approach to intermediate variables in causal inference. *Tech. Rep.*, Duke University, USA.

Sfer, A. M. (2005). *Randomization and Causality.* Ph.D. Thesis, Universidad Nacional de Tucumán, Argentina.

Slavković, A. B. and Fienberg, S. E. (2010). Algebraic geometry of 2 × 2 contingency tables. *Algebraic and Geometric Methods in Statistics* (P. Gibilisco, *et al.*, eds.). Cambridge: Cambridge University Press, 67–85.

Zhang, J. L., Rubin, D. B., Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. *Modelling and Evaluating Treatment Effects in Econometrics* (D. L. Millimet, J. A. Smith and E. J. Vytlacil, eds.). Amsterdam: Elsevier, 117–145.

Zhang, J. L., Rubin, D. B., Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104**, 166–176.