Nonparametric Bounds on Causal Effects from Partial Compliance Data

Alexander Balke and Judea Pearl*
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
balke@cs.ucla.edu
judea@cs.ucla.edu

Abstract

Experimental studies in which treatment assignment is random but subject compliance is imperfect may be susceptible to bias; the actual effect of the treatment may deviate appreciably from the mean difference between treated and untreated subjects. This paper establishes universal formulas that can be used to bound the actual treatment effect in any experiment for which compliance data is available and in which the assignment influences the response only through the treatment given. Using a linear programming analysis, we present formulas that provide the tightest bounds that can be inferred on the average treatment effect, given an empirical distribution of assignments, treatments, and responses. The application of these results is demonstrated on data that relates cholesterol levels to cholestyramine treatment ([Lipid Research Clinic Program 84]).

KEY WORDS: Causal model, Latent variable, Linear programming, Rubin's model.

*The research was partially supported by Air Force grant #AFOSR 90 0136, NSF grant #IRI-9200918, Northrop Micro grant #92-123, and Rockwell Micro grant #92-122. Alexander Balke was supported by the Fannie and John Hertz Foundation. An earlier version of Sections 2 and 3 appeared in the *Proceedings of the 49th Session of the International Statistical Institute: Invited papers*, Florence, Italy, August 1993. This work benefitted from discussions with Joshua Angrist, David Chickering, Thomas Ferguson, David Galles, Guido Imbens, Paul Rosenbaum, and Donald Rubin. The authors thank Bradley Efron for providing us with the data used in Section 5.

1 INTRODUCTION

Consider an experimental study where random assignment has taken place but compliance is not perfect (i.e., the treatment received differs from that assigned). It is well known that under such conditions a bias may be introduced, in the sense that the true causal effect of the treatment may deviate substantially from the causal effect computed by simply comparing subjects receiving the treatment with those not receiving the treatment. Because the subjects who did not comply with the assignment may be precisely those who would have responded adversely (positively) to the treatment, the actual effect of the treatment, when applied uniformly to the population, might be substantially less (more) effective than the study reveals.

In an attempt to avert this bias, economists have devised correctional formulas based on a model called "instrumental variables" ([Bowden and Turkington 84]) which, in general, do not hold outside the linear regression model. A recent analysis by [Efron and Feldman 91] departs from the linear regression model, but still makes restrictive commitments to a particular mode of interaction between compliance and response. [Manski 90] has derived nonparametric bounds on treatment effects under rather general conditions which, unfortunately, do not reflect the unique, two-stage, partly randomized process characteristic of studies with imperfect compliance. [Holland 88] has given a general formulation of the problem (which he called "encouragement design") in terms of Rubin's model of causal effect and has outlined its relation to path analysis and structural equations models. [Angrist et al. 93], also invoking Rubin's model, have identified a set of assumptions under which the "Instrumental Variable" formula is valid for certain subpopulations. These subpopulations cannot be identified from empirical observation alone, and the need remains to devise alternative, assumption-free formulas for assessing the effect of treatment over the population as a whole. In this paper, we derive bounds on the average treatment effect that rely solely on observed quantities and are universal, that is, valid no matter what model actually governs the interactions between compliance and response.

The paper is organized as follows. In Section 2 we formulate our assumptions using a graphical model, and we present a simple, closed-form formula for bounding treatment effect, given partial compliance data. Section 3 reformulates the problem using a potential-response model — closely related to Rubin's ([Rubin 74]) model of counterfactuals — in which the states of some variables correspond to hypothetical response policies of subjects in the population. In Section 4, using a linear programming analysis of the potential-response model, we refine the bounds presented in Section 3 and present the results in tabular form. For any experimental distribution of assignments, treatments, and responses, the tables provide the tightest bounds that can be inferred on the actual treatment effect. In Section 5, the application of these bounds is demonstrated on the [Lipid Research Clinic Program 84] data, which relate cholestyramine treatments to cholesterol levels. In Section 6, we show by hypothetical example how the bounds derived in Section 4 can significantly improve upon the bounds from Section 2. In Section 7, we introduce special assumptions that might further restrict subject behavior, and for each such assumption we derive the corresponding bounds on the treatment effect.

2 BOUNDS DERIVED FROM A GRAPHICAL MODEL

The canonical partial-compliance setting can be graphically modeled as shown in Figure 1.

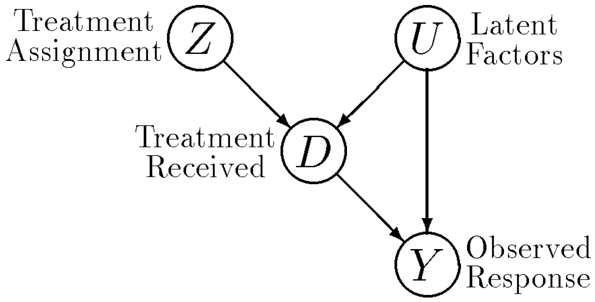

Figure 1: *Graphical representation of causal dependencies in a randomized clinical trial with partial compliance.*

We assume that Z, D, and Y are observed binary variables where Z represents the (randomized) treatment assignment, D is the treatment actually received, and Y is the observed response. U represents all factors, both observed and unobserved, that may influence the outcome Y and the treatment D. To facilitate the notation, we let z, d, and y represent, respectively, the values taken by the variables Z, D, and Y, with the following interpretation:
$z \in \{z_0, z_1\}$, z_1 asserts that treatment has been assigned (z_0, its negation);
$d \in \{d_0, d_1\}$, d_1 asserts that treatment has been administered (d_0, its negation); and
$y \in \{y_0, y_1\}$, y_1 asserts a positive observed response (y_0, its negation).
The domain of U remains unspecified and may, in general, combine the spaces of several random variables, both discrete and continuous.

The graphical model reflects two assumptions of independence:

1. The treatment assignment does not influence Y directly, but only through the actual treatment D, that is,

$$Z \perp\!\!\!\perp Y \mid \{D, U\} \tag{1}$$

   In practice, any direct effect Z might have on Y would be adjusted for through the use of a placebo.

2. Z and U are marginally independent, that is, $Z \perp\!\!\!\perp U$. This independence is partly ensured through the randomization of Z, which rules out a common cause for both Z and U. The absence of a direct path from Z to U represents the assumption that a person's disposition to comply with or deviate from a given assignment is not in itself affected by the assignment; any such effect can be viewed as part of the disposition.

These assumptions impose on the joint distribution[1] the decomposition

$$P(y,d,z,u) \;=\; P(y|d,u)\;P(d|z,u)\;P(z)\;P(u) \tag{2}$$

which, of course, cannot be observed directly because U is a latent variable. However, the marginal distribution $P(y,d,z)$ and, in particular, the conditional distributions $P(y,d|z), z \in \{z_0, z_1\}$, are observed, and the challenge is to assess the causal effect of D on Y from these distributions.[2]

In addition to the independence assumption above, the graphical model of Figure 1 reflects claims about the behavior of the population under external interventions. In particular, it reflects the assumption that $P(y|d,u)$ is a stable quantity: the probability that an individual with characteristics $U = u$ given treatment $D = d$ will respond with $Y = y$ remains the same, regardless of how the treatment was selected — be it by choice or by policy. Therefore, if we wish to predict the distribution of Y under a condition where the treatment D is applied uniformly to the population, we should calculate

$$P(Y=y|D=d \text{ applied uniformly}) = E[P(y|d,u)] \tag{3}$$

where E stands for the expectation taken over u.

Likewise, if we are interested in estimating the average *change* in Y due to treatment, we define the average *causal effect*, $\mathrm{ACE}(D \rightarrow Y)$ ([Holland 88]), as

$$\mathrm{ACE}(D \rightarrow Y) \;=\; E[P(y_1|d_1,u) - P(y_1|d_0,u)] \tag{4}$$

The task of causal inference is then to estimate or bound the expectation in Eq. (4), given the observed probabilities $P(y,d|z_0)$ and $P(y,d|z_1)$.

For uniformity of notation, we can define, in an analogous way, the average causal effects of the assignment Z, $\mathrm{ACE}(Z \rightarrow Y)$ and $\mathrm{ACE}(Z \rightarrow D)$. However, since Z is chosen at random, averaging over u is superfluous, and these two quantities can be obtained from the observed distribution:

$$\mathrm{ACE}(Z \rightarrow D) \;=\; P(d_1|z_1) - P(d_1|z_0) \tag{5}$$

$$\mathrm{ACE}(Z \rightarrow Y) \;=\; P(y_1|z_1) - P(y_1|z_0) \tag{6}$$

After a few algebraic manipulations (see [Pearl 93]), Eq. (4) yields an alternative expression for $\mathrm{ACE}(D \rightarrow Y)$:

$$\mathrm{ACE}(D \rightarrow Y) = E\left[\frac{P(y_1|z_1,u) - P(y_1|z_0,u)}{P(d_1|z_1,u) - P(d_1|z_0,u)}\right] \tag{7}$$

If we think of u as an index characterizing the experimental units (i.e., the subjects), the result is simple and intuitive. It says that for each individual unit u, the indirect

[1]Only the expectation over U will enter our analysis, hence we take the liberty of denoting the prior distribution of U by $P(u)$, even though U may consist of continuous variables.

[2]In practice, of course, only a finite sample of $P(y,d|z)$ will be observed, but since our task is one of identification, not estimation, we make the large-sample assumption and consider $P(y,d|z)$ as given.

causal effect along the chain $Z \rightarrow D \rightarrow Y$ is equal to the product of the individual causal effects along the two links of the chain. If all units were to exhibit the same difference in compliance probabilities, $P(d_1|z_1, u) - P(d_1|z_0, u)$, we would have obtained the celebrated "Instrumental Variable" formula

$$\text{ACE}(D \rightarrow Y) = \frac{\text{ACE}(Z \rightarrow Y)}{\text{ACE}(Z \rightarrow D)} = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1) - P(d_1|z_0)} \tag{8}$$

which says that the causal effect $\text{ACE}(Z \rightarrow Y)$ associated with the intent-to-treat needs to be adjusted upward, through division by the degree of compliance $\text{ACE}(Z \rightarrow D)$. This ratio formula is indeed valid in linear regression models, under which it was derived by social scientists and econometrician as far back as 1940 ([Angrist et al. 93, Holland 88]). In general, however, the quantities on the right-hand side of Eq. (7) cannot be observed directly (only in expectation), and $\text{ACE}(D \rightarrow Y)$ can become as low as zero or even negative. Still, when almost perfect compliance is observed, the unknown quantities $P(y|d, u)$, $P(d|z, u)$, and $P(u)$ do not have the freedom to render $\text{ACE}(D \rightarrow Y)$ substantially different from $\text{ACE}(Z \rightarrow Y)$, and informative bounds can then be obtained on the actual causal effect of the treatment.

Further analysis of Eqs. (2) and (3) ([Pearl 93]) yields the following bounds for the two terms on the right-hand side of Eq. (4):

$$\max[P(y_1, d_1|z_1);\ P(y_1, d_1, |z_0)] \leq E[P(y_1|d_1, u)] \leq 1 - \max[P(y_0, d_1|z_0);\ P(y_0, d_1|z_1)] \tag{9}$$

$$\max[P(y_1, d_0|z_0);\ P(y_1, d_0, |z_1)] \leq E[P(y_1|d_0, u)] \leq 1 - \max[P(y_0, d_0|z_0);\ P(y_0, d_0|z_1)] \tag{10}$$

Choosing appropriate terms to bound the difference $E[P(y_1|d_1, u)] - E[P(y_1|d_0, u)]$, we obtain lower and upper bounds on the causal effect of D on Y:

$$P(y_1, d_1|z_1) + P(y_0, d_0|z_0) - 1 \leq \text{ACE}(D \rightarrow Y) \leq 1 - P(y_0, d_1|z_1) - P(y_1, d_0|z_0) \tag{11}$$

or, alternatively,

$$\begin{aligned} \text{ACE}(D \rightarrow Y) &\geq \text{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\ \text{ACE}(D \rightarrow Y) &\leq \text{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0) \end{aligned} \tag{12}$$

Due to its simplicity and wide range of applicability, we will call the bounds of Eq. (12) the *natural bounds* (three other less intuitive expressions for the upper and lower bounds may be inferred from Eqs. (9) and (10), but these will not be presented here because they will be rederived in Section 4). The natural bounds guarantee that the causal effect of the actual treatment cannot exceed that of the intent-to-treat by more than the sum of two measurable quantities, $P(y_1, d_0|z_1) + P(y_0, d_1|z_0)$; they also guarantee that the causal effect of treatment cannot drop below that of the intent-to-treat by more than the sum of two other measurable quantities, $P(y_0, d_0|z_1) + P(y_1, d_1|z_0)$. The width of the natural bound, not surprisingly, is given by the rate of defection, $P(d_1|z_0) + P(d_0|z_1)$. While the bounds in Eqs. (9) and (10) are sharp, the ones in Eq. (12) can be substantially improved using linear programming (see Section 4), albeit at the expense of formal elegance.

Before continuing to the more refined model of Section 3, we should point out that the structural model of Figure 1 imposes definite constraints on the observed distributions $P(y,d|z_0)$ and $P(y,d|z_1)$. The constraints, obtained directly from Eq. (2), are

$$\begin{aligned}
P(y_1,d_1|z_1) &\leq 1-P(y_0,d_1|z_0)\\
P(y_1,d_1|z_0) &\leq 1-P(y_0,d_1|z_1)\\
P(y_1,d_0|z_1) &\leq 1-P(y_0,d_0|z_0)\\
P(y_1,d_0|z_0) &\leq 1-P(y_0,d_0|z_1)
\end{aligned}$$

These constraints constitute necessary and sufficient conditions for a marginal probability distribution $P(y,d,z)$ to be generated by the structure of Figure 1, and therefore they may serve as an operational test for the compatibility of that structure with the observed data.

3 THE POTENTIAL-RESPONSE MODEL

A powerful feature of the graphical model discussed so far is its capacity for producing meaningful results while keeping the latent variable U totally unspecified; U may be finite or unbounded, discrete or continuous, ordered as well as unstructured. Although this generality has the advantage of freeing the analyst from making a commitment to a particular parametric model, it may turn into an inconvenience when finer mathematical details are needed. Fortunately, it can be shown ([Pearl 93]) that it is always possible to replace the latent variable U, no matter how complex, by two discrete and finite variables, one representing tendencies to comply with assignment, the other representing tendencies to respond to treatment.

Figure 2 depicts a structure, equivalent to that of Figure 1, in which the latent variables R and R' have only four states each: $r \in \{r_0,r_1,r_2,r_3\}$ and $r' \in \{r'_0,r'_1,r'_2,r'_3\}$. It was shown in [Pearl 93] that every experimental outcome modeled by the graphical structure of Figure 1 can also fit into the finite-variable structure of Figure 2 and, moreover, that the states of the variables R and R' correspond to the *potential-response* vectors in Rubin's model of causal effects ([Rubin 74, Rosenbaum and Rubin 83]), as defined below.

D is a deterministic function of the variables Z and R:

$$d = F_D(z,r) = \begin{cases}
d_0 & \text{if} \quad r=r_0 \\
\\
d_0 & \text{if} \quad r=r_1 \quad z=z_0 \\
d_1 & \text{if} \quad r=r_1 \quad z=z_1 \\
\\
d_1 & \text{if} \quad r=r_2 \quad z=z_0 \\
d_0 & \text{if} \quad r=r_2 \quad z=z_1 \\
\\
d_1 & \text{if} \quad r=r_3
\end{cases}$$

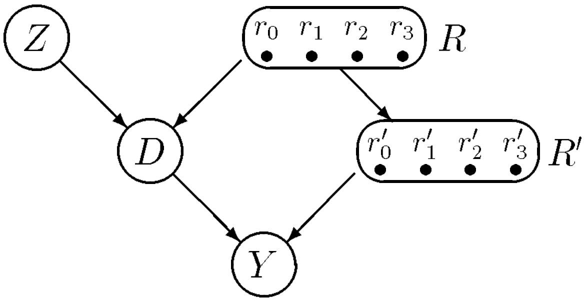

Figure 2: *A structure equivalent to that of Figure 1 but employing two latent variables, R and R', with four states each.*

Similarly, Y is a deterministic function of D and R':

$$y = F_Y(d, r') = \begin{cases} y_0 & \text{if} \quad r' = r'_0 \\ \\ y_0 & \text{if} \quad r' = r'_1 \quad d = d_0 \\ y_1 & \text{if} \quad r' = r'_1 \quad d = d_1 \\ \\ y_1 & \text{if} \quad r' = r'_2 \quad d = d_0 \\ y_0 & \text{if} \quad r' = r'_2 \quad d = d_1 \\ \\ y_1 & \text{if} \quad r' = r'_3 \end{cases} \tag{13}$$

The correspondence between the states of variables R and R' and the potential response vectors in the Rubin's model is rather transparent: each state corresponds to a counterfactual statement specifying how a unit in the population (e.g., a person) would have reacted to any given input. For example, r_1 represents units with perfect compliance, while r_2 represents units with perfect defiance. Similarly, r'_1 represents units with perfect response to treatment, while r'_0 represents units with no response ($y = y_0$) regardless of treatment. The counterfactual variables Y_1 and Y_0 usually invoked in Rubin's model can be obtained from R' as follows:

$$\begin{aligned} Y_1 = \quad \{Y \text{ if } D = d_1\} \quad &= \begin{cases} 1 & \text{if } R' = r'_1 \text{ or } R' = r'_3 \\ 0 & \text{otherwise} \end{cases} \\ Y_0 = \quad \{Y \text{ if } D = d_0\} \quad &= \begin{cases} 1 & \text{if } R' = r'_2 \text{ or } R' = r'_3 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In general, treatment response and compliance attitudes may not be independent, hence the arrow $R \rightarrow R'$ in Figure 2. The joint distribution over $R \times R'$ requires 15 independent parameters, and these parameters are sufficient for specifying the model of Figure 2, $P(y, d, z, r, r') = P(y|d, r')P(d|r, z)P(z)P(r, r')$, because Y and D stand in functional relation to their parents in the graph. The causal effect of the treatment can now be obtained directly from Eqs. (3) and (13), giving

$$\begin{aligned} P(y_1|D = d_1 \text{ applied uniformly}) &= P(r' = r'_1) + P(r' = r'_3) \\ P(y_1|D = d_0 \text{ applied uniformly}) &= P(r' = r'_2) + P(r' = r'_3) \end{aligned}$$

and

$$\mathrm{ACE}(D \rightarrow Y) = P(r' = r'_1) - P(r' = r'_2) \tag{14}$$

The computational advantage of the potential-response model is twofold. First, lower bounds on $\mathrm{ACE}(D \rightarrow Y)$ can now be produced by minimizing a linear function over a 15-dimensional vector space, rather than by dealing with the unspecified domain of U. Second, the constraints that the data $P(y, d|z_0)$ and $P(y, d|z_1)$ induce on the parameters of $P(r, r')$ are linear, compared with the non-convex constraints induced on the parameters $P(d|z, u)$ and $P(y|d, u)$ in the graphical structure of Figure 1. This enables the use of linear programming techniques to obtain tighter bounds on the causal effect $\mathrm{ACE}(D \rightarrow Y)$; such bounds are much harder to obtain in a model where U remains unspecified.

4 TIGHT BOUNDS ON TREATMENT EFFECTS

4.1 Linear programming formulation

In this section we will explicate the relationship between the parameters of the observed distribution $P(y, d|z)$ and the parameters of the joint distribution $P(r, r')$ of the potential-response functions. This will lead directly to the linear constraints needed for minimizing/maximizing $\mathrm{ACE}(D \rightarrow Y)$ given the observation $P(y, d|z)$.

The conditional distribution $P(y, d|z)$ over the observable variables is fully specified by eight parameters, which will be notated as follows:

$$\begin{aligned}
p_{00.0} &= P(y_0, d_0|z_0) \\
p_{01.0} &= P(y_0, d_1|z_0) \\
p_{10.0} &= P(y_1, d_0|z_0) \\
p_{11.0} &= P(y_1, d_1|z_0) \\
\\
p_{00.1} &= P(y_0, d_0|z_1) \\
p_{01.1} &= P(y_0, d_1|z_1) \\
p_{10.1} &= P(y_1, d_0|z_1) \\
p_{11.1} &= P(y_1, d_1|z_1)
\end{aligned}$$

The probabilistic constraints

$$\sum_{n=00}^{11} p_{n.0} = 1 \tag{15}$$

$$\sum_{n=00}^{11} p_{n.1} = 1 \tag{16}$$

further imply that $\vec{p} = (p_{00.0}, p_{01.0}, p_{10.0}, p_{11.0}, p_{00.1}, p_{01.1}, p_{10.1}, p_{11.1})$ can be specified by a point in six-dimensional space. This space will be referred to as P. Eqs. (5) and

(6) may be rewritten in terms of these parameters as

$$\mathrm{ACE}(Z \rightarrow D) \;=\; p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} \tag{17}$$
$$\mathrm{ACE}(Z \rightarrow Y) \;=\; p_{11.1} + p_{10.1} - p_{11.0} - p_{10.0} \tag{18}$$

The joint probability over $R \times R'$, $P(r, r')$, has 16 parameters and completely specifies the population under study. These parameters will be notated as

$$q_{jk} \;=\; P(r = r_j, r' = r'_k)$$

where $j, k \in \{0, 1, 2, 3\}$. The probabilistic constraint

$$\sum_{j=0}^{3} \sum_{k=0}^{3} q_{jk} \;=\; 1$$

implies that $\vec{q} = (q_{00}, q_{01}, q_{02}, q_{03}, q_{10}, q_{11}, q_{12}, q_{13}, q_{20}, q_{21}, q_{22}, q_{23}, q_{30}, q_{31}, q_{32}, q_{33})$ specifies a point in 15-dimensional space. This space will be referred to as Q.

Eq. (14) can now be rewritten as a linear combination of the Q parameters:

$$\mathrm{ACE}(D \rightarrow Y) \;=\; q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32} \tag{19}$$

Given some point $\vec{q}$ in Q space, there is a direct linear transformation to the corresponding point $\vec{p}$ in the observation space P:

$$\begin{aligned}
p_{00.0} &= q_{00} + q_{01} + q_{10} + q_{11} \\
p_{01.0} &= q_{20} + q_{22} + q_{30} + q_{32} \\
p_{10.0} &= q_{02} + q_{03} + q_{12} + q_{13} \\
p_{11.0} &= q_{21} + q_{23} + q_{31} + q_{33} \\
\\
p_{00.1} &= q_{00} + q_{01} + q_{20} + q_{21} \\
p_{01.1} &= q_{10} + q_{12} + q_{30} + q_{32} \\
p_{10.1} &= q_{02} + q_{03} + q_{22} + q_{23} \\
p_{11.1} &= q_{11} + q_{13} + q_{31} + q_{33}
\end{aligned}$$

which will sometimes be written in matrix form, $\vec{p} = \bar{P}\vec{q}$.

Given a point $\vec{p}$ in P space, the strict lower bound on $\mathrm{ACE}(D \rightarrow Y)$ can be determined by solving the following linear programming problem:

Minimize: $q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32}$

Subject to:

$$\begin{aligned}
\sum_{j=0}^{3} \sum_{k=0}^{3} q_{jk} &= 1 \\
\bar{P}\vec{q} &= \vec{p} \\
q_{jk} &\geq 0 \text{ for } j, k \in \{0, 1, 2, 3\}
\end{aligned} \tag{20}$$

4.2 Symbolic solutions to the linear programming problem

Given an observed point $\vec{p}$ in P space, $L_{D\to Y}(\vec{p})$ and $U_{D\to Y}(\vec{p})$, respectively, will represent the strict lower and upper bounds on $\mathrm{ACE}(D \to Y)$ associated with $\vec{p}$. More precisely,

$$L_{D\to Y}(\vec{p}) = \min_{\vec{q} \text{ s.t. } \vec{p}=\bar{P}\vec{q}} \mathrm{ACE}(D \to Y) \tag{21}$$

$$U_{D\to Y}(\vec{p}) = \max_{\vec{q} \text{ s.t. } \vec{p}=\bar{P}\vec{q}} \mathrm{ACE}(D \to Y) \tag{22}$$

where Eq. (19) gives $\mathrm{ACE}(D \to Y)$ in terms of $\vec{q}$.

For every given point $\vec{p}$, the optimization above can be executed using the Simplex Tableau algorithm (see [Davis and McKeown 81]), which yields a pair of numerical values for $L_{D\to Y}(\vec{p})$ and $U_{D\to Y}(\vec{p})$. Fortunately, the size of the problem permits a symbolic solution to be obtained by tracking the conditions that lead to the various decisions in the Simplex Tableau algorithm. This procedure generates a decision tree with 34 leaf nodes, each containing a symbolic solution to $L_{D\to Y}(\vec{p})$. By taking the union of the conditions leading to leaf nodes with identical expressions, eight distinct formulas were obtained, each conditioned by a conjunction of inequalities as presented in Table 1. The first entry in the table corresponds to the natural lower bound of Eq. (11).

Conditions	$L_{D\to Y}(\vec{p})$
$p_{11.1} \geq p_{11.0}$ $p_{01.1} + p_{10.1} \geq p_{01.0}$ $p_{00.0} \geq p_{00.1}$ $p_{01.0} + p_{10.0} \geq p_{10.1}$	$p_{11.1} + p_{00.0} - 1$
$p_{11.0} \geq p_{11.1}$ $p_{01.0} + p_{10.0} \geq p_{01.1}$ $p_{00.1} \geq p_{00.0}$ $p_{01.1} + p_{10.1} \geq p_{10.0}$	$p_{11.0} + p_{00.1} - 1$
$p_{11.0} \geq p_{11.1} + p_{10.1}$ $p_{01.1} \geq p_{01.0} + p_{10.0}$	$p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0}$
$p_{11.1} \geq p_{11.0} + p_{10.0}$ $p_{01.0} \geq p_{01.1} + p_{10.1}$	$p_{11.1} - p_{01.1} - p_{10.1} - p_{11.0} - p_{10.0}$
$p_{11.0} + p_{10.0} \geq p_{11.1} \geq p_{11.0}$ $p_{01.0} + p_{00.0} \geq p_{00.1} \geq p_{00.0}$	$-p_{01.1} - p_{10.1}$
$p_{11.1} + p_{10.1} \geq p_{11.0} \geq p_{11.1}$ $p_{01.1} + p_{00.1} \geq p_{00.0} \geq p_{00.1}$	$-p_{01.0} - p_{10.0}$
$p_{10.0} \geq p_{01.1} + p_{10.1}$ $p_{00.1} \geq p_{01.0} + p_{00.0}$	$p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0}$
$p_{10.1} \geq p_{01.0} + p_{10.0}$ $p_{00.0} \geq p_{01.1} + p_{00.1}$	$p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1}$

Table 1: *Lower bounds on* $\mathrm{ACE}(D \to Y)$ *given a point* $\vec{p}$ *in the observation space* P.

A more convenient representation of this table is obtained by noting that $L_{D\to Y}(\vec{p})$ is simply the maximum of the eight expressions in the right-hand column of the table.

Thus,

$$L_{D\to Y}(\vec{p}) \quad = \quad \max \left\{ \begin{array}{c} p_{11.1} + p_{00.0} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0} \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} \\ -p_{01.1} - p_{10.1} \\ -p_{01.0} - p_{10.0} \\ p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0} \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} \end{array} \right\} \qquad (23)$$

The upper bound can be derived in similar fashion, by maximizing rather than minimizing the objective function. However, instead of duplicating the maximizing exercise, we can take advantage of the fact that the solution of the lower bound problem gives us the maximally negative causal effects, which correspond to the upper bound on the causal effect if we switch the label on the observed treatment response variable Y and take the additive inverse of each solution. The results of this exercise are shown in Table 2. The first entry in the table corresponds to the natural upper bound of Eq. (11).

Conditions	$U_{D\to Y}(\vec{p})$
$p_{01.1} \geq p_{01.0}$ $p_{11.1} + p_{00.1} \geq p_{11.0}$ $p_{10.0} \geq p_{10.1}$ $p_{11.0} + p_{00.0} \geq p_{00.1}$	$1 - p_{01.1} - p_{10.0}$
$p_{01.0} \geq p_{01.1}$ $p_{11.0} + p_{00.0} \geq p_{11.1}$ $p_{10.1} \geq p_{10.0}$ $p_{11.1} + p_{00.1} \geq p_{00.0}$	$1 - p_{01.0} - p_{10.1}$
$p_{01.0} \geq p_{01.1} + p_{00.1}$ $p_{11.1} \geq p_{11.0} + p_{00.0}$	$-p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0}$
$p_{01.1} \geq p_{01.0} + p_{00.0}$ $p_{11.0} \geq p_{11.1} + p_{00.1}$	$-p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0}$
$p_{01.0} + p_{00.0} \geq p_{01.1} \geq p_{01.0}$ $p_{11.0} + p_{10.0} \geq p_{10.1} \geq p_{10.0}$	$p_{11.1} + p_{00.1}$
$p_{01.1} + p_{00.1} \geq p_{01.0} \geq p_{01.1}$ $p_{11.1} + p_{10.1} \geq p_{10.0} \geq p_{10.1}$	$p_{11.0} + p_{00.0}$
$p_{00.0} \geq p_{11.1} + p_{00.1}$ $p_{10.1} \geq p_{11.0} + p_{10.0}$	$-p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0}$
$p_{00.1} \geq p_{11.0} + p_{00.0}$ $p_{10.0} \geq p_{11.1} + p_{10.1}$	$-p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1}$

Table 2: *Upper bounds on* ACE$(D \to Y)$ *given a point* $\vec{p}$ *in observation space* P.

As for the lower bound case, we can show that the conditions on each formula imply

$$U_{D\to Y}(\vec{p}) = \min\left\{\begin{array}{c} 1 - p_{01.1} - p_{10.0} \\ 1 - p_{01.0} - p_{10.1} \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} \\ -p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0} \\ p_{11.1} + p_{00.1} \\ p_{11.0} + p_{00.0} \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} \\ -p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1} \end{array}\right\} \quad (24)$$

4.3 The positive-effects convention

To simplify the presentation of the bounds found in the last subsection, we first choose a notational system in which assignment to treatment does not reduce the probability of treatment usage ($D = d_1$) and of positive response ($Y = y_1$). From Eqs. (5) and (6), these conditions can be written as

$$\begin{aligned} \mathrm{ACE}(Z \to D) &\geq 0 \\ \mathrm{ACE}(Z \to Y) &\geq 0 \end{aligned}$$

or, alternatively,

$$\begin{aligned} p_{01.1} + p_{11.1} &\geq p_{01.0} + p_{11.0} \\ p_{10.1} + p_{11.1} &\geq p_{10.0} + p_{11.0} \end{aligned}$$

The conjunction of these two inequalities will be referred to as the *condition of positive effects*. This constraint may be imposed without loss of generality, because the labels of the variables' values can always be swapped in such a way that the inequalities are satisfied: if $\mathrm{ACE}(Z \to D) < 0$, we swap d_0 and d_1; if $\mathrm{ACE}(Z \to Y) < 0$, we swap y_0 and y_1.

In a notational system where the condition of positive effects holds, the lower and upper bounds on the treatment effect can be simplified to read

$$L_{D\to Y}(\vec{p}) = \max\left\{\begin{array}{c} p_{11.1} + p_{00.0} - 1 \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} \\ -p_{01.1} - p_{10.1} \\ -p_{01.0} - p_{10.0} \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} \end{array}\right\} \quad (25)$$

and

$$U_{D\to Y}(\vec{p}) = \min\left\{\begin{array}{c} 1 - p_{01.1} - p_{10.0} \\ 1 - p_{01.0} - p_{10.1} \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} \\ p_{11.1} + p_{00.1} \\ p_{11.0} + p_{00.0} \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} \end{array}\right\} \quad (26)$$

respectively.

4.4 Graphical presentation of the bounds

When compliance is perfect (i.e., ACE($Z \to D$) = 1), we expect the causal effect of the treatment to coincide with the causal effect of the intent-to-treat, that is,

$$\text{ACE}(D \to Y) = \text{ACE}(Z \to Y) \quad \text{if} \quad \text{ACE}(Z \to D) = 1$$

Similarly, if the conditions for the Instrumental Variable formula (Eq. (8)) are satisfied (e.g., linear models), we expect ACE($D \to Y$) to be determined solely by ACE($Z \to Y$) and ACE($Z \to D$). In general, however, the latter two parameters will not be sufficient to determine ACE($D \to Y$) uniquely; nevertheless, they can be used to determine the range within which ACE($D \to Y$) may fall.

Figure 3 plots $L_{D \to Y}(\vec{p})$ and $U_{D \to Y}(\vec{p})$ given ACE($Z \to D$) and ACE($Z \to Y$). The range of ACE($D \to Y$) is quite wide, and is given by the simple formula:

$$\text{ACE}(Z \to Y) + \text{ACE}(Z \to D) - 1 \leq \text{ACE}(D \to Y) \leq 1 - |\text{ACE}(Z \to D) - \text{ACE}(Z \to Y)| \tag{27}$$

An interesting point is that plotting the natural bounds given by Eq. (12) as a function of ACE($Z \to D$) and ACE($Z \to Y$) gives us precisely the same results as shown in Figure 3.

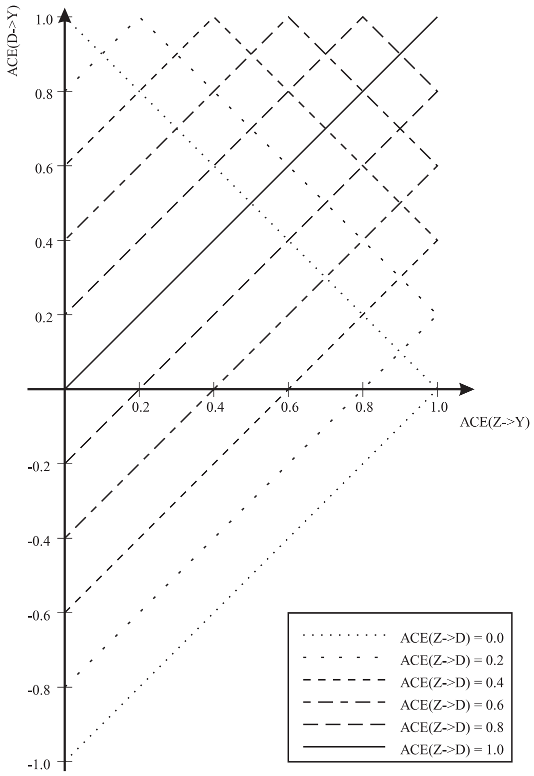

Figure 3: *Bounds on* ACE($D \to Y$) *plotted against* ACE($Z \to Y$) *and* ACE($Z \to D$).

Note that the bounds $L_{D \to Y}(\vec{p})$ and $U_{D \to Y}(\vec{p})$ for a particular point $\vec{p}$ in P space may be much tighter than the bounds shown in Figure 3 as functions of ACE($Z \to D$) and ACE($Z \to Y$) evaluated at $\vec{p}$. This will be demonstrated by example in Section 5.

5 EXAMPLES

At this point it is worth summarizing by example how the bounds of Eqs. (23) and (24) can be used to provide meaningful information about causal effects.

Consider the Lipid Research Clinics Coronary Primary Prevention Trial data (see [Lipid Research Clinic Program 84] for an extended description of the clinical trial). A portion of this data consisting of 337 subjects was analyzed in [Efron and Feldman 91] using a model that incorporated subject compliance as an explanatory variable; this same data set is the focus of our analysis. A population of subjects was assembled and two preliminary cholesterol measurements were obtained: one prior to a suggested low-cholesterol diet (continuous variable C_{I1}); and one following the diet period (C_{I2}). The initial cholesterol level (C_I) was taken as a weighted average of these two measures: $C_I = 0.25C_{I1} + 0.75C_{I2}$. The subjects were randomized into two treatment groups; in the first group all subjects were prescribed cholestyramine (z_1), while the subjects in the other group were prescribed a placebo (z_0). During several years of treatment, each subject's cholesterol level was measured multiple times, and the average of these measurements was used as the post-treatment cholesterol level (continuous variable C_F). The compliance of each subject was determined by tracking the quantity of prescribed dosage consumed (continuous variable B).

In order to apply our analysis to this study, the continuous data obtained in the [Lipid Research Clinic Program 84] study must be transformed to binary variables representing *treatment assignment* (Z), *received treatment* (D), and *treatment response* (Y). The following transformation accomplishes this by thresholding dosage consumption and change in cholesterol level:

$$d = \begin{cases} d_0 & \text{if } z = z_0 \text{ or } b < 50 \\ d_1 & \text{if } z = z_1 \text{ and } b \geq 50 \end{cases} \tag{28}$$

$$y = \begin{cases} y_0 & \text{if } c_I - c_F < 28 \\ y_1 & \text{if } c_I - c_F \geq 28 \end{cases} \tag{29}$$

This transformation reflects the assumption that a subject does not receive cholestyramine if not assigned to the cholestyramine treatment group, namely, $P(y_0, d_1|z_0) = 0$ and $P(y_1, d_1|z_0) = 0$. The threshold for dosage consumption in Eq. (28) was selected as roughly the midpoint between minimum and maximum consumption, while the threshold for cholesterol level reduction in Eq. (29) was selected at 28 units.

If the data samples are interpreted according to Eqs. (28) and (29), then the computed distribution over (Z, D, Y) results in the following point in P space[3]:

$$\begin{aligned} p_{00.0} = P(y_0, d_0|z_0) &= 0.919 \\ p_{01.0} = P(y_0, d_1|z_0) &= 0.000 \\ p_{10.0} = P(y_1, d_0|z_0) &= 0.081 \\ p_{11.0} = P(y_1, d_1|z_0) &= 0.000 \end{aligned}$$

[3]We make the large-sample assumption and take the sample frequencies as representing $P(y, d|z)$.

$$
\begin{aligned}
p_{00.1} = P(y_0, d_0|z_1) &= 0.315 \\
p_{01.1} = P(y_0, d_1|z_1) &= 0.139 \\
p_{10.1} = P(y_1, d_0|z_1) &= 0.073 \\
p_{11.1} = P(y_1, d_1|z_1) &= 0.473
\end{aligned}
$$

By first computing the causal effects of the intent-to-treat,

$$
\begin{aligned}
\text{ACE}(Z \to D) &= p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} = 0.612 \\
\text{ACE}(Z \to Y) &= p_{11.1} + p_{10.1} - p_{11.0} - p_{10.0} = 0.465
\end{aligned}
\tag{30}
$$

we can verify that the condition of positive effects is satisfied. This justifies the use of Eqs. (25) and (26) for evaluating the strict lower and upper bounds on ACE($D \to Y$). By computing the quantities required for Eq. (25), we obtain

$$
L_{D\to Y}(\vec{p}) = \max \left\{ \begin{array}{rcr}
p_{11.1} + p_{00.0} - 1 &=& 0.392 \\
p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} &=& 0.180 \\
-p_{01.1} - p_{10.1} &=& -0.212 \\
-p_{01.0} - p_{10.0} &=& -0.081 \\
p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} &=& 0.384
\end{array} \right\}
\tag{31}
$$

Those needed for Eq. (26) give us

$$
U_{D\to Y}(\vec{p}) = \min \left\{ \begin{array}{rcr}
1 - p_{01.1} - p_{10.0} &=& 0.780 \\
1 - p_{01.0} - p_{10.1} &=& 0.927 \\
-p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} &=& 1.373 \\
p_{11.1} + p_{00.1} &=& 0.788 \\
p_{11.0} + p_{00.0} &=& 0.919 \\
-p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} &=& 0.796
\end{array} \right\}
\tag{32}
$$

Accordingly, we conclude that the treatment causal effect lies in the range

$$
0.392 \le \text{ACE}(D \to Y) \le 0.780 \tag{33}
$$

which is rather remarkable; the experimenter can categorically state that when applied uniformly to the population, the treatment is guaranteed to improve by at least 39.2% the probability of reducing the level of cholesterol by at least 28 points. This guarantee does not rest on any assumed model. Unfortunately, these results cannot be translated directly into a useful policy statement for treating people with high cholesterol, because the [Lipid Research Clinic Program 84] data were obtained for continuous level of dosage consumed (D), while our analysis is restricted to binary D. To infer the behavior of the population under uniform consumption at a specific level of dosage, a model with a continuous (or at least 3-level) treatment must be studied.

Note that the bounds in Eq. (33) are equal to the natural bounds given by Eq. (12):

$$
\begin{aligned}
\text{ACE}(D \to Y) &\ge 0.465 - 0.073 - 0.000 = 0.392 \\
\text{ACE}(D \to Y) &\le 0.465 + 0.315 + 0.000 = 0.780
\end{aligned}
$$

It is interesting to note that "naive" comparison of subjects in and out of the treatment group would predict, in this case, the value of

$$P(y_1|d_1) - P(y_1|d_0) = 0.662 \tag{34}$$

which demonstrates the potential inaccuracy in using the mean difference for evaluating ACE($D \rightarrow Y$).

If ACE($Z \rightarrow D$) and ACE($Z \rightarrow Y$) are the only quantities measured, then the following bounds on ACE($D \rightarrow Y$) can be computed by substituting the values from Eq. (30) into Eq. (27):

$$0.077 \leq \text{ACE}(D \rightarrow Y) \leq 0.853 \tag{35}$$

As noted in Section 4.4, these bounds are much wider than those obtained in Eq. (33), which utilized the full information given by $P(y, d|z)$.

6 TIGHTNESS OF THE NATURAL BOUND

Although the example above shows no improvement over the natural bounds, the next (hypothetical) example will show that in certain cases the natural bounds can be improved upon significantly. Consider the following point in P space:

$$\begin{aligned}
p_{00.0} = P(y_0, d_0|z_0) &= 0.55 \\
p_{01.0} = P(y_0, d_1|z_0) &= 0.45 \\
p_{10.0} = P(y_1, d_0|z_0) &= 0.00 \\
p_{11.0} = P(y_1, d_1|z_0) &= 0.00 \\
\\
p_{00.1} = P(y_0, d_0|z_1) &= 0.45 \\
p_{01.1} = P(y_0, d_1|z_1) &= 0.00 \\
p_{10.1} = P(y_1, d_0|z_1) &= 0.00 \\
p_{11.1} = P(y_1, d_1|z_1) &= 0.55
\end{aligned}$$

Substitution of these parameters into Eq. (12) results in the natural bounds

$$0.10 \leq \text{ACE}(D \rightarrow Y) \leq 0.55$$

while the bounds resulting from the application of Eqs. (23) and (24) collapse to

$$0.55 \leq \text{ACE}(D \rightarrow Y) \leq 0.55$$

Obviously, when our goal is the assessment of the treatment causal effect, the bounds obtained through linear programming can be much more informative.

Interestingly, a precise determination of ACE($D \rightarrow Y$) is feasible even though the compliance is low:

$$\text{ACE}(Z \rightarrow D) = 0.10$$

Intuitively, one would expect that if most subjects ignore their treatment assignment, the results of the study would be suspect. This intuition is partially supported by Figure 3, which shows that the feasible range of ACE($D \rightarrow Y$) tends to widen as ACE($Z \rightarrow D$) decreases. Nevertheless, the idiosyncratic features of the data in this example permit us to determine precisely the causal effect. These features also allow us to precisely determine the distribution of subjects in the population, in terms of the subjects' compliance and response characteristics.

This example is just one in a whole class of points in P space:

$$\begin{aligned}
p_{00.0} &= \frac{1}{2}(1+x) \\
p_{01.0} &= \frac{1}{2}(1-x) \\
p_{10.0} &= 0 \\
p_{11.0} &= 0 \\
p_{00.1} &= \frac{1}{2}(1-x) \\
p_{01.1} &= 0 \\
p_{10.1} &= 0 \\
p_{11.1} &= \frac{1}{2}(1+x)
\end{aligned}$$

for $0 \leq x < 1$, where the upper and lower bounds coincide

$$\begin{aligned}
L_{D\rightarrow Y}(\vec{p}) &= \frac{1}{2}(1+x) \\
U_{D\rightarrow Y}(\vec{p}) &= \frac{1}{2}(1+x)
\end{aligned}$$

while the natural bounds give

$$x \leq \text{ACE}(D \rightarrow Y) \leq \frac{1}{2}(1+x)$$

The tight lower and upper bounds are equal because the observed points in this class can only be modelled by a single point in Q space.

Each point in this class represents a rather odd population, one in which subjects fall into only one of two behaviors with the following distribution:

$$\begin{aligned}
q_{11} = P(r = r_1, r' = r'_1) &= \frac{1}{2}(1+x) \\
q_{20} = P(r = r_2, r' = r'_0) &= \frac{1}{2}(1-x)
\end{aligned}$$

The first behavior is characterized by perfect compliance with the assignment along with a perfect response pattern to the treatment received ($y = y_1$ if and only if $d = d_1$). The second behavior is characterized by perfect defiance of the assignment (the subject always chooses the treatment that is the opposite of the one assigned) along with a total inability to respond positively to either treatment. The strong and strange interactions between the compliance and response behaviors implied by these data would be very uncharacteristic of most subject populations.

In this section, we have shown that, in general, the natural bounds given by Eq. (12) may not always be as tight as the bounds given by Eqs. (23) and (24). In the next section, however, we will demonstrate that the natural bounds are tight in two important subspaces of P: when the data reveal treatment sufficiency (conditional independence between treatment assignment and treatment response given treatment received), and when it is reasonable to assume that subjects are *non-defiant*.

7 INCORPORATING ADDITIONAL ASSUMPTIONS

In this section we will examine the impact that various assumptions have on the bounds for $\mathrm{ACE}(D \to Y)$ and the constraints that they place on the observed parameters. The main assumptions to be discussed here are:

- treatment sufficiency (conditional independence of treatment assignment and observed response given treatment received);
- treatment sufficiency together with structural stability; and
- no perfectly defiant subjects.

7.1 Treatment sufficiency

This subsection examines whether the presence of conditional independence $Z \perp\!\!\!\perp Y|D$ in the data simplifies the formulas for the bounds on $\mathrm{ACE}(D \to Y)$. In other words, are any of the expressions within the minimization/maximization of Eqs. (25) and (26) eliminated? The following theorem provides the answer to this question.

Theorem 7.1 *If the observed distribution $P(y, d|z)$ satisfies $Z \perp\!\!\!\perp Y|D$ and the condition of positive effects, then the natural bounds on* $\mathrm{ACE}(D \to Y)$

$$
\begin{aligned}
\mathrm{ACE}(D \to Y) &\geq \mathrm{ACE}(Z \to Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\
\mathrm{ACE}(D \to Y) &\leq \mathrm{ACE}(Z \to Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0)
\end{aligned}
$$

are tight.

Proof:

We will show that a set of constraints implied by $Z \perp\!\!\!\perp Y|D$ and the condition of positive effects are only mutually consistent with those conditions in Tables 1 and 2 corresponding to the natural bounds (the topmost entries).

First, assume that $\vec{p}$ is strictly positive.

By definition, $Z \perp\!\!\!\perp Y|D$ if and only if

$$P(y|d, z_0) = P(y|d, z_1)$$

for all y and d such that $P(d|z_0) > 0$ and $P(d|z_1) > 0$. This may be written:

$$\begin{aligned}\frac{p_{10.0}}{p_{00.0} + p_{10.0}} &= \frac{p_{10.1}}{p_{00.1} + p_{10.1}} \\ \frac{p_{11.0}}{p_{01.0} + p_{11.0}} &= \frac{p_{11.1}}{p_{01.1} + p_{11.1}}\end{aligned}$$

or, equivalently,

$$\begin{aligned} p_{00.1} &= S p_{00.0} \\ p_{10.1} &= S p_{10.0} \\ p_{01.0} &= T p_{01.1} \\ p_{11.0} &= T p_{11.1} \end{aligned} \tag{36}$$

where S and T represent the ratios

$$\begin{aligned} S &= \frac{p_{00.1}}{p_{00.0}} = \frac{p_{10.1}}{p_{10.0}} \\ T &= \frac{p_{01.0}}{p_{01.1}} = \frac{p_{11.0}}{p_{11.1}} \end{aligned}$$

From the condition of positive effects,

$$p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} \geq 0$$

which, from Eq. (36), may be rewritten

$$(1 - T)(p_{11.1} + p_{01.1}) \geq 0 \tag{37}$$

This implies that $T \leq 1$.

Likewise, we may use the equalities in Eq. (36) to rewrite the probabilistic constraints given by Eqs. (15) and (16):

$$\begin{aligned} p_{00.0} + T p_{01.1} + p_{10.0} + T p_{11.1} &= 1 \\ S p_{00.0} + p_{01.1} + S p_{10.0} + p_{11.1} &= 1 \end{aligned}$$

Taking the difference of these two equations gives

$$(1 - S)(p_{00.0} + p_{10.0}) = (1 - T)(p_{01.1} + p_{11.1}) \tag{38}$$

$T \leq 1$ then implies that $S \leq 1$.

Applying these bounds on S and T to Eq. (36) results in the constraints

$$\begin{aligned} p_{00.0} &\geq p_{00.1} \\ p_{10.0} &\geq p_{10.1} \\ p_{01.1} &\geq p_{01.0} \\ p_{11.1} &\geq p_{11.0} \end{aligned}$$

which, when conjoined with the conditions in Tables 1 and 2, reveal that the only applicable bounds on ACE($D \rightarrow Y$) under the assumption of positive effects and conditional independence are the natural bounds:

$$
\begin{aligned}
L_{D\rightarrow Y}(\vec{p}) &= p_{11.1} + p_{00.0} - 1 = & \text{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\
U_{D\rightarrow Y}(\vec{p}) &= 1 - p_{01.1} - p_{10.0} = & \text{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0)
\end{aligned}
$$

When p is not strictly positive, we can proceed through a similar exercise on a case-by-case basis and obtain identical results. We omit this part of the proof.

□

Figure 4 shows how the conditional independence tightens the lower bounds shown in Figure 3 when the only information known about the observed distribution is ACE($Z \rightarrow D$) and ACE($Z \rightarrow Y$).

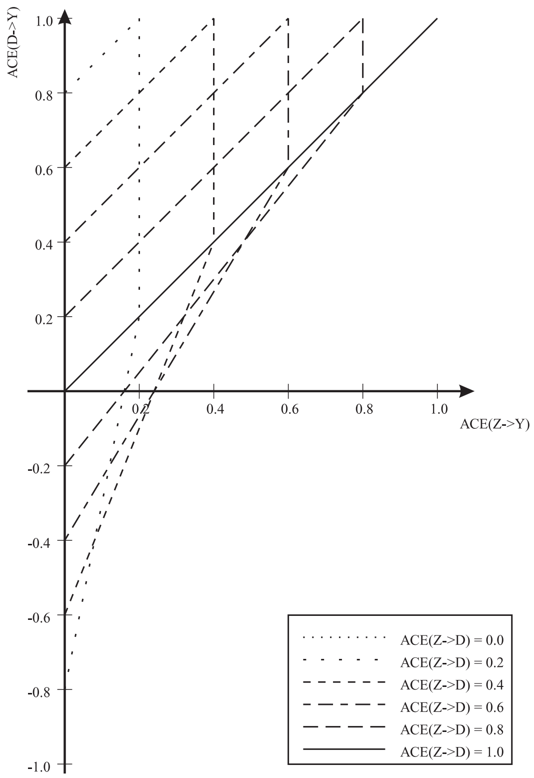

Figure 4: *Bounds on* ACE($D \rightarrow Y$) *plotted against* ACE($Z \rightarrow Y$) *and* ACE($Z \rightarrow D$), *given that* $Z \perp\!\!\!\perp Y|D$.

7.2 Treatment sufficiency with structural stability

Where treatment sufficiency holds under a variety of experimental conditions, it is reasonable to assume that it is not caused by incidental equality of parameters, but rather by structural constraints. This notion of structural stability is indeed the pivotal assumption behind the causal inference methods of [Pearl and Verma 91, Spirtes et al. 91], namely, that every conditional independence shown in the data must be logically implied by the decomposition of the joint probability distribution given by Eq. (2) as dictated by the graph structure. If this assumption holds, then the data are *DAG-isomorphic* to the graph structure, and all independence relations may then be tested by using the *d-separation* criterion ([Pearl 88]).

Theorem 7.2 *If an observed distribution $P(y, d|z)$ is structurally stable and satisfies $Y \perp\!\!\!\perp Z|D$ and $Y \not\perp\!\!\!\perp Z$, then*

$$\text{ACE}(D \to Y) = P(y_1|d_1) - P(y_1|d_0) \tag{39}$$

Proof:

The antecedent of the theorem implies that Z and Y must be d-separated given D in the graph structure for which the data is DAG-isomorphic. Applying the d-separation criterion to the graphical structure of Figure 1, we find that, given D, Z and Y are dependent via the path, $Z - D - U - Y$. The only way to remove this dependency is to eliminate one of the following edges: $Z \to D$, $U \to D$, or $U \to Y$. The assumption that Z and D are marginally dependent prevents the elimination of $Z \to D$; therefore, the antecedent of the theorem can only be satisfied if at least one of the edges $U \to D$ or $U \to Y$ is eliminated.

First, assume that $U \to Y$ is eliminated from the graph structure. In this case, $P(y|d, u) = P(y|d)$, which, when substituted into Eq. (4), results in

$$\text{ACE}(D \to Y) \;=\; P(y_1|d_1) - P(y_1|d_0)$$

Next, assume that $U \to D$ is eliminated from the graph structure. In this case, we note that $P(u) = P(u|d)$, allowing the following transformations of Eq. (4):

$$\begin{aligned}
\text{ACE}(D \to Y) &= \sum_u [P(u)P(y_1|d_1, u) - P(u)P(y_1|d_0, u)] \\
&= \sum_u [P(u|d_1)P(y_1|d_1, u) - P(u|d_0)P(y_1|d_0, u)] \\
&= \sum_u [P(y_1, u|d_1) - P(y_1, u|d_0)] \\
&= P(y_1|d_1) - P(y_1|d_0)
\end{aligned}$$

□

Notice that the combination of structural stability and treatment sufficiency subsumes the assumption of Eq. (1); $Z \perp\!\!\!\perp Y|\{D, U\}$ is no longer an assumption but is implied by $Z \perp\!\!\!\perp Y|D$, because, for any set of variables S, $Z \perp\!\!\!\perp Y|S$ cannot hold if there is a direct arc from Z to Y. Therefore, when structural stability holds, finding a variable Z' satisfying $Z' \perp\!\!\!\perp Y|D$ and $Z' \not\perp\!\!\!\perp Y$ permits us to dispose of the randomized assignment altogether and infer causal effects (using Eq. (39)) in purely observational studies. Discovering a Z' which satisfies these relationships may be viewed as uncovering a randomized experiment that is conducted by Nature itself, and this is the basis of the "virtual control" condition discussed in [Pearl and Verma 91].

7.3 Non-defiance

In Section 3, a subject was characterized as *perfectly defiant* if under either treatment assignment the subject fails to comply with the assignment ($d = d_1$ if and only if $z = z_0$). In terms of the potential-response model, this behavior is specified by $R = r_2$. An example of a situation in which perfectly defiant subjects might be possible is a study that involves observation of draft status (Z) and military service (D) ([Angrist et al. 93]). In this scenario, there could conceivably be subjects who despise authority and so, if drafted, would evade service and, if not drafted, would volunteer for service.

Alternatively, there are situations in which perfectly defiant behavior would be improbable:

- when subjects do not know exactly what the two treatment options (z_0 and z_1) are; hence, it is beyond their means to defy both treatment assignments.
- when subjects know what the two treatment options are but do not know which treatment they have been assigned (the procedures for receiving the assigned treatments are identical, as in the use of placebo).
- when subjects know what both treatments are and know which treatment they have been assigned but do not have access to both treatments; therefore, it is beyond their means to obtain the opposite treatment under either assignment.

Drug studies often are very likely to fit one of these situations, especially since a placebo is usually used as the alternative treatment to the medication under study, so subjects cannot easily determine which treatment they have been assigned.

Based on the applicability suggested above, we will define the assumption of *non-defiance* as stating that there are no perfectly defiant subjects in a study. This assumption is expressed by the constraint $P(r = r_2) = 0$, or $q_{2j} = 0$ for $j = 0, \ldots, 3$. Non-defiance together with the condition of positive effects is equivalent to the assumption of "monotonicity" analyzed by [Angrist et al. 93], which translates to the restriction: either $P(r = r_2) = 0$ or $P(r = r_1) = 0$. Because the assumption of non-defiance imposes restrictions on the unobserved parameters in Q space, it carries the potential of improving the bounds on $\mathrm{ACE}(D \rightarrow Y)$ beyond those of Eqs. (21) and (22). The following theorem refutes this possibility.

Theorem 7.3 *If all subjects in a population are non-defiant, then the natural bounds on* $\mathrm{ACE}(D \rightarrow Y)$*,*

$$\begin{aligned}
\mathrm{ACE}(D \rightarrow Y) &\geq \mathrm{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\
\mathrm{ACE}(D \rightarrow Y) &\leq \mathrm{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0)
\end{aligned}$$

are tight.

This theorem may be proven by reapplying the Simplex Tableau algorithm to the optimization problem given by Eq. (20) with the constraints $q_{2j} = 0$ for $j = 0, \ldots, 3$. Exactly as before, we obtain symbolic solutions for the upper and lower bounds by tracking the conditions that lead to various decisions in the Simplex Tableau algorithm. This procedure results in a single expression each for the lower and upper bounds; these expressions are identical to the natural bounds given by Eq. (12).

It is important to understand that the non-defiance assumption (as well as that of treatment sufficiency) does not widen the bounds of Eqs. (21) and (22) to the natural bounds, but instead restricts the observation space P to a region where the natural bounds are the only applicable bounds. Consequently, the assumption of non-defiance is partly observable; if $P(y, d|z)$ does not satisfy the following constraints implied by non-defiance

$$\begin{aligned} p_{00.0} &\geq p_{00.1} \\ p_{01.1} &\geq p_{01.0} \\ p_{10.0} &\geq p_{10.1} \\ p_{11.1} &\geq p_{11.0} \end{aligned}$$

then the assumption of non-defiance does not hold. To summarize, the assumption of non-defiance provides no benefits over the unconditional bounds given by Eqs. (23) and (24); however, it narrows the space of observation so as to render the natural bounds of Eq. (12) realizable.

7.4 Local average-treatment effect

While this paper focuses primarily on predicting the average treatment effect over an entire population, there are cases where one would be interested in treatment effects averaged over a subpopulation of special characteristics. [Angrist et al. 93] have found that, under the assumption of non-defiance, the treatment effect averaged over the subpopulation of perfectly complying individuals, $\mathrm{ACE}_c(D \to Y)$, can be identified and is given by the Instrumental Variable formula

$$\mathrm{ACE}_c(D \to Y) = \frac{\mathrm{ACE}(Z \to Y)}{\mathrm{ACE}(Z \to D)} = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1) - P(d_1|z_0)} \tag{40}$$

In other words, Eq. (40) gives the correct treatment effect for those individuals whose participation in the treatment D comes as a consequence of the encouragement Z.

This can be verified by noting that a compliant subpopulation is characterized by the condition $R = r_1$; thus

$$\begin{aligned} \mathrm{ACE}_c(D \to Y) &= P(y_1|d_1, r_1) - P(y_1|d_0, r_1) \\ &= P(r_1'|r_1) - P(r_2'|r_1) \\ &= \frac{P(r_1, r_1') - P(r_1, r_2')}{P(r_1)} \\ &= \frac{q_{11} - q_{12}}{q_{10} + q_{11} + q_{12} + q_{13}} \end{aligned}$$

This last expression coincides with the Instrumental Variable formula above under the condition of non-defiance, namely, $P(r_2) = 0$, or $q_{2j} = 0$ for $j = 0, \ldots, 3$.

It is worth noting that the subpopulation of perfectly complying individuals is not, in general, identifiable, because the condition $R = r_1$ cannot be determined from the triplet (y, d, z). Nevertheless, the behavior of this subpopulation may be of interest to analysts, as it reveals the treatment effect under ideal conditions, free of noncompliance side effects. Bounds on the behavior of other subpopulations of interest can be obtained by methods similar to those in Section 4.

8 CONCLUSIONS

This paper provides formulas that allow analysts to make categorical statements about causal effects in the context of studies where subjects are only partially compliant. These formulas, expressed in terms of the distribution over observed variables (treatment assignment, treatment received, and observed response), represent strict upper and lower bounds for the average causal effect of the treatment on the population. These bounds are applicable to all studies where the assignment itself only affects the observed response via the treatment actually received, regardless of any interaction that might take place between the treatment received and the observed response. Aside from this assumption, the results do not rest on any particular model of compliance behavior.

We believe that the results presented here could be particularly helpful in quasi-experimental studies, that is, studies in which randomized mandated treatments are either unfeasible or undesirable and randomized encouragements are instituted instead ([Holland 88]). For example, in evaluating the efficacy of a social program, the randomized instrument can be advertisement, incentives, or eligibility, letting subjects make the final choice of participation. The bounds established through Eqs. (23) and (24) reveal that such studies, despite the indirectness of the randomized instrument, can yield valuable information on the average causal effect of the treatment on the population.

Some topics that will receive attention in future work include the analysis of multi-level and continuous treatments, the maximum-likelihood estimation technique for finite samples, and the analysis of hypothetical queries.

References

[Angrist et al. 93] J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. Technical Report No. 136, Department of Economics, Harvard University, Cambridge, MA, June 1993.

[Bowden and Turkington 84] Roger J. Bowden and Darrell A. Turkington. *Instrumental Variables*. Cambridge University Press, Cambridge, MA, 1984.

[Davis and McKeown 81] K. Roscoe Davis and Patrick G. McKeown. *Quantitative Models for Management*. Kent Publishing Company, Boston, MA, 1981.

[Efron and Feldman 91] B. Efron and D. Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–26, March 1991.

[Holland 88] Paul W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington, DC, 1988.

[Lipid Research Clinic Program 84] Lipid Research Clinic Program. The Lipid Research Clinics Coronary Primary Prevention Trial results, parts I and II. *Journal of the American Medical Association*, 251(3):351–374, January 1984.

[Manski 90] Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323, May 1990.

[Pearl 88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman Publishers, San Mateo, CA, 1988.

[Pearl 93] Judea Pearl. Aspects of graphical models connected with causality. Technical Report R-195-LL, Cognitive Systems Laboratory, UCLA, June 1993. Submitted to Biometrika (June 1993). Short version in *Proceedings of the 49th Session of the International Statistical Institute: Invited papers*, Florence, Italy, August 1993, Tome LV, Book 1, pp. 391–401.

[Pearl and Verma 91] Judea Pearl and Thomas Verma. A theory of inferred causation. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference.* Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[Rosenbaum and Rubin 83] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[Rubin 74] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[Spirtes et al. 91] Peter Spirtes, Clark Glymour, and Richard Scheines. From probability to causality. *Philosophical Studies*, 64(1):1–36, October 1991.