Causal Inference from Indirect Experiments

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
Phone: (310) 825-3243
Fax: (310) 825-2273
judea@cs.ucla.edu

Abstract

Indirect experiments are studies in which randomized control is replaced by randomized *encouragement*, that is, subjects are encouraged, rather than forced to receive treatment programs. The purpose of this paper is to bring to the attention of experimental researchers simple mathematical results that enable us to assess, from indirect experiments, the strength with which causal influences operate among variables of interest. The results reveal that despite the laxity of the encouraging instrument, indirect experimentation can yield significant and sometimes accurate information on the impact of a program on the population as a whole, as well as on the particular individuals who participated in the program.

Keywords: Causal reasoning, treatment evaluation, noncompliance, graphical models

1 Introduction

Standard experimental studies in the biological, medical, and behavioral sciences invariably invoke the instrument of randomized control, that is, subjects are assigned at random to various groups (or treatments or programs) and the mean differences between participants in different groups are regarded as measures of the efficacies of the associated programs. *Indirect experiments* are studies in which randomized control is either infeasible or undesirable, and randomized *encouragement* instituted instead, that is, subjects are still assigned at random to various groups, but members of each group are encouraged, rather than forced, to participate in the program associated with the group; it is up to the individuals to select among the programs.

Recently, use of randomization in social and medical experimentation has been questioned. The objections seem to fall into three major categories:

1. Perfect control is hard to achieve or ascertain. Studies in which treatment is assumed to be randomized may be marred by uncontrolled imperfect compliance. For example, subjects experiencing adverse reactions to an experimental drug may decide to reduce the assigned dosage. Such imperfect compliance introduces appreciable bias into the conclusions that researchers draw from data, and this bias cannot be corrected unless detailed models of compliance are constructed [7].

2. Denying subjects assigned to certain control groups the benefits of the best available treatment has moral and legal ramifications. For example, it is difficult to justify placebo programs in AIDS research because those patients assigned to the placebo group would be denied access to potentially life saving treatment [13].

3. Randomization, by its very presence, may influence participation as well as behavior [9]. For example, eligible candidates may be wary of applying to a school once they discover that it deliberately randomizes its admission criteria. Likewise, as Kramer and Shapiro [10] note, subjects in drug trials were less likely to participate in randomized trials than in nonexperimental studies, even when the treatments were equally nonthreatening.

Altogether, researchers are finally being forced to acknowledge that mandated randomization may undermine the reliability of experimental evidence and that experimentation with human subjects should include an element of *self selection*.

This paper concerns the drawing of inferences from studies in which subjects are indeed given final choice of program, while randomization is confined to an indirect instrument that merely encourages or discourages participation in the various programs. For example, in evaluating the efficacy of a given training program, notices of eligibility may be sent to a randomly selected group of students or, alternatively, eligible candidates may be selected at random to receive scholarships for participating in the program. Similarly, in drug trials, subjects may be given randomly chosen advice on recommended dosage level, yet the final choice of dosage will be determined by the subjects to fit their individual needs.

The question we attempt to answer in this investigation is whether such indirect randomization can provide sufficient information to allow accurate assessment of the intrinsic merit of a program, as would be measured, for example, if the program were to be extended and mandated uniformly to the population. The analysis presented shows that, given a minimal set of assumptions, such inferences are indeed possible, albeit in the form of *bounds*, rather than precise point estimates, for the causal effect of the program or treatment. These bounds can be used by the analyst to guarantee that the causal impact of a given program must be higher than one measurable quantity and lower than another.

Our most crucial assumption is that, for any given person, the encouraging instrument influences the treatment chosen by that person but has no effect on how that person would respond to the treatment chosen. The second assumption, one which is always made in experimental studies, is that subjects respond to treatment independently of one other. Other than these two assumptions, our model places no constraints on how tendencies to respond to treatments may interact with choices among treatments.

2 Problem Statement

The basic experimental setting associated with indirect experimentation is shown in Figure 1. To focus the discussion, we will consider a prototypical clinical trial with partial compliance although, in general, the model applies to any study in which a randomized instrument encourages subjects to choose one program over another.

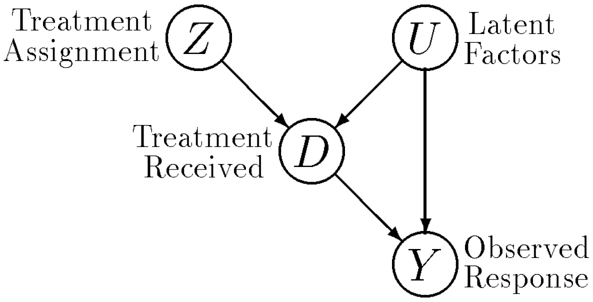

Figure 1: *Graphical representation of causal dependencies in a randomized clinical trial with partial compliance.*

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 influence the way a subject responds to treatments; hence, an arrow is drawn from U to Y. The arrow from U to D denotes that the U factors may also influence the subject's choice of treatment D. For example, subjects who could benefit most from the treatment may be precisely those who decide, perhaps due to adverse initial reaction, to stop the treatment and switch to the control group. Thus, although D is shown to depend directly on Z and U, this dependence may represent a complex decision process standing between the assignment (Z) and the actual 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:

1. The assigned treatment Z does not influence Y directly but rather through the actual treatment D. 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, as ensured through the randomization of Z, which rules out a common cause for both Z and U.

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

$$P(y, d, z, u) \;=\; P(y|d, u)\; P(d|z, u)\; P(z)\; P(u) \tag{1}$$

which, of course, cannot be observed directly because U is unobserved. However, the marginal distribution $P(y, d, z)$ and, in particular, the conditional distribution $P(y, d|z), z \in \{z_0, z_1\}$, are observed,[2] and the challenge is to assess from these distributions the average *change* in Y due to treatment.

If $P(y|d,\ u)$ gives the probability that an individual with characteristic $U = u$ will respond with $Y = y$ to a treatment $D = d$, then taking the average over u gives the average response if the treatment d is applied uniformly to the population. Therefore, the *average causal effect* of D on Y, α, is defined as the difference

$$\alpha \;=\; E[P(y_1|d_1, u) - P(y_1|d_0, u)] \tag{2}$$

when E stands for the expectation taken over u.

[1] Only the expectation over U will enter our analysis, hence we take the liberty of denoting the 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.

When compliance is perfect, D and U are independent, $P(y|d,\ u)$ can be written $P(y,\ d|u)/P(d|u) = P(y,\ d|u)/P(d)$, and $E[P(y|d,\ u)]$ reduces to the observed conditional probability $P(y|d)$. Thus, α would be measured by the observed mean difference between treated and untreated subjects,

$$\Delta(Y|D) \quad = \quad P(y_1|d_1) \ - \ P(y_1|d_0) \tag{3}$$

However, when compliance is not perfect, high values of $\Delta(Y)$ may correspond to low or even negative values of α. The discrepancy between α and $\Delta(Y)$ is known as *confounding bias*, which, clearly, cannot be eliminated by taking a larger sample size.

To circumvent this bias, analysts usually advocate the use of *intent-to-treat analysis*, which compares assignment groups regardless of the treatment actually received. In other words, instead of estimating α, we settle for

$$\Delta(Y|Z) = P(y_1|z_1) - P(y_1|z_0)$$

which represents the causal effect of the encouragement instrument Z on Y, since Z is randomized. Estimates based on this analysis are free of confounding bias as long as the experimental conditions perfectly mimic the conditions prevailing in the eventual usage of the treatment. In particular, the experiment should mimic subjects' incentives for receiving each treatment. In situations where field incentives are more compelling than experimental incentives, treatment effectiveness is determined by α, and estimates based on $\Delta(Y|Z)$ can be extremely misleading. For example, imagine a study in which a drug has an adverse effect on a large segment of the population, and that only those members of the segment who drop from the treatment arm recover. The $\Delta(Y|Z)$ measure will attribute these cases of recovery to the drug since they are part of the intent-to-treat arm, while in reality these cases have recovered by avoiding the treatment. The formulas reported in this paper should enable the analyst to determine the extent to which estimates based on intent-to-treat analysis deviate from the actual treatment effect. This information should be useful for assessing whether efforts to ensure population compliance have the potential to increase or decrease the overall benefit of the treatment.

Much of the statistical literature assumes that α is the parameter of interest, because it predicts the impact of applying the treatment uniformly (or randomly) over the population. However, if a policy maker is not interested in introducing new treatment policies but rather in deciding whether to maintain or terminate an existing program under its current incentive system, then the parameter of interest should measure the impact of the treatment *on the treated*:

$$\alpha^* = E[P(y_1|d_1, u) - P(y_1|d_0, u)|D = d_1] \tag{4}$$

namely, the change of the mean response of the treated subjects compared to the mean response of these same subjects had they not been treated [9]. The formulas reported in this paper provide assessments for both α and α^*.

3 Summary of Results

Analysis shows [19, 12, 14] that the expression for α (Eq. (2)) can be bounded by two simple formulas, each made up of observed parameters of $P(y, d|z)$ (see Appendix):

$$
\begin{aligned}
\alpha &\geq P(y_1|z_1) - P(y_1|z_0) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\
\alpha &\leq P(y_1|z_1) - P(y_1|z_0) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0)
\end{aligned} \tag{5}
$$

Due to their simplicity and wide range of applicability, the bounds of Eq. (5) were named the *natural bounds* [2]. The natural bounds guarantee that the causal effect of the actual treatment cannot be lower than that of the encouragement 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 the treatment cannot exceed that of the encouragement 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 bounds, not surprisingly, is given by the rate of noncompliance, $P(d_1|z_0) + P(d_0|z_1)$.

This width can be narrowed further using linear programming [2], which shows that, even under conditions of imperfect compliance, some experimental data (i.e., $P(x, y|z)$) can permit the precise evaluation of α. These narrower bounds, which are in fact the narrowest possible, are given by the following inequalities:

$$
E[P(y_1|d, u] \geq \max \left\{ \begin{array}{l}
P(y_1, d|z_0) \\
P(y_1, d|z_1) \\
P(y_1|z_0) - P(y_1, \overline{d}|z_1) - P(y_0, d|z_1) \\
P(y_1|z_1) - P(y_1, \overline{d}|z_0) - P(y_0, d|z_0)
\end{array} \right\} \triangleq L(d)
$$

$$
E[P(y_1|d, u] \leq \min \left\{ \begin{array}{l}
P(y_0|z_1) + P(y_1, \overline{d}|z_1) \\
P(y_0|z_0) + P(y_1, \overline{d}|z_0) \\
P(y_1|z_1) + P(y_0, \overline{d}|z_0) + P(y_1, d|z_0) \\
P(y_1|z_0) + P(y_0, \overline{d}|z_1) + P(y_1, d|z_1)
\end{array} \right\} \triangleq U(d) \tag{6}
$$

where d and its complement $\overline{d}$ range over $\{d_0, d_1\}$. These inequalities induce the following bounds on α:

$$
U(d_1) - L(d_0) \geq \alpha \geq L(d_1) - U(d_0) \tag{7}
$$

where $U(d)$ and $L(d)$ are defined in Eq. (6).

Although more complicated than the natural bounds of Eq. (5), these expressions are nevertheless easy to assess once we have the frequency data in the eight cells of $P(y, d|z)$. It can also be shown that the natural bounds are optimal where we can safely assume that no subject is *contrarian*, that is, that no subject would consistently choose a treatment arm contrary to the one assigned.

The analysis also shows that α^* can be assessed with greater accuracy than α. More remarkably, under conditions of *no intrusion* (namely, $P(d_1|z_0) = 0$, as in most clinical trials), α^* can be identified precisely [1].

The bounds governing α^* are (see Appendix)

$$\begin{aligned}
\alpha^* &\geq \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} - \frac{P(y_0, d_1|z_0)}{P(d_1)} \\
\alpha^* &\leq \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} + \frac{P(y_1, d_1|z_0)}{P(d_1)}
\end{aligned} \tag{8}$$

Clearly, in situations where treatment may only be obtained by those encouraged (by assignment), α^* is identifiable and is given by

$$\alpha^* = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1)} \tag{9}$$

if $P(d_1|z_0) = 0$. Unlike the α measure, α^* is not an intrinsic property of the treatment, as it varies with the encouraging instrument. As noted in Section 2, the significance in the α^* measure emerges primarily in studies where it is desired to evaluate the efficacy of an *existing* program on its current participants. In such studies, under the assumption that encouragement is randomized, α^* is given simply by the mean response difference between the encouraged and nonencouraged populations, divided by the rate of participation $P(d_1|z_1)$.

4 Example 1: The Effects of Cholestyramine

To demonstrate by example how the bounds for α can be used to provide meaningful information about causal effects, consider the Lipid Research Clinics Coronary Primary Prevention Trial data (see [11]). A portion of this data consisting of 337 subjects was analyzed in [7] and is the focus of this example. Subjects were randomized into two treatments groups of roughly equal size; in one 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 (a continuous quantity).

In order to apply the bounds of Eq. (5) to this study, the continuous data is first transformed, using thresholds, to binary variables representing treatment assignment (Z), received treatment (D), and treatment response (Y). The threshold for dosage consumption was selected as roughly the midpoint between minimum and maximum consumption, while the threshold for cholesterol level reduction was set at 28 units.

The data samples after thresholding give rise to the following eight probabilities[3]:

$$
\begin{array}{rclrcl}
P(y_0, d_0|z_0) & = & 0.919 & P(y_0, d_0|z_1) & = & 0.315 \\
P(y_0, d_1|z_0) & = & 0.000 & P(y_0, d_1|z_1) & = & 0.139 \\
P(y_1, d_0|z_0) & = & 0.081 & P(y_1, d_0|z_1) & = & 0.073 \\
P(y_1, d_1|z_0) & = & 0.000 & P(y_1, d_1|z_1) & = & 0.473
\end{array}
$$

This data represents a compliance rate of

$$P(d_1|z_1) = 0.139 + 0.473 = 0.61$$

a mean difference (using $P(z_1) = 0.50$) of

$$\Delta(Y|D) = P(y_1|d_1) - p(y_1|d_0) = \frac{0.473}{0.473 + 0.139} - \frac{0.073 + 0.081}{1 + 0.315 + 0.073} = 0.662$$

and an encouragement effect of

$$\Delta(Y|Z) = P(y_1|z_1) - P(y_1|z_0) = 0.073 + 0.473 - 0.081 = 0.465$$

According to Eq. (5), α can be bounded by

$$
\begin{array}{rcl}
\alpha & \geq & 0.465 - 0.073 - 0.000 = 0.392 \\
\alpha & \leq & 0.465 + 0.315 + 0.000 = 0.780
\end{array}
$$

These are remarkably informative bounds: although 38.8% of the subjects deviated from their treatment protocol, 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 28 points or more. This guarantee is purely mathematical and does not rest on any assumed model of subject behavior.

The impact of treatment "on the treated" is equally revealing. Using Eq. (9), α^* can be evaluated precisely (since $P(d_1|z_0) = 0$):

$$\alpha^* = \frac{0.465}{0.610} = 0.762$$

In other words, those subjects who stayed in the program are much better off than they would have been if not treated: the treatment can be credited with reducing cholesterol levels by at least 28 units in 76.2% of these subjects.

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

5 Example 2: The Paradoxical Episode of Dr. Pearson

5.1 The Story

The school board decided to hire an expert statistician, Dr. Pearson, to examine the question of whether there is a real difference between the educational effectiveness of the town's two high schools, A and B. After extensive discussions with all parties involved, Dr. Pearson decided to resolve the issue by conducting a clean randomized experiment. He obtained the list of all students about to enroll in next year's freshman classes and selected at random two groups of 100 students each. He then went to the homes of all 200 students and persuaded parents to enroll their children in school A or school B, depending on the group the child happened to be in. Dr. Pearson then went back to his work at the university, after arranging to come back in the spring to analyze the test scores of the two groups.

When he returned, Dr. Pearson found that something had gone terribly wrong immediately after he left town. Some kids found their assigned school too boring, others complained about school being too hard, and still others were plain contrarians. Altogether, only 50% of those assigned to school A actually remained in school A after the first week of classes, and exactly the same had happened with those assigned to school B. Dr. Pearson was totally depressed. While it is true that the test scores showed a marked performance difference between the two groups – 50% of the students who went to school B and none of those attending school A managed to pass the state exam – he knew very well that the study was worthless by all statistical standards.

As Dr. Pearson disgustedly prepared to quit town, his young assistant, Alex, asked to have one more look at the data. "What is there to look at?" snarled Dr. Pearson, "I can tell a classic case of noncompliance when I see one! Even when only 10% of the subjects switch groups, I have a hell of a time convincing my colleagues that the study is worth a nickel. Can you imagine how they would react if I were to show them data with a 50% crossover rate? They would say that exactly those students who would have performed better in school A decided to switch over to school B and that those who switched from school B to school A were precisely the students who would not have performed well no matter where they went. You know how convoluted statisticians can be".

Surprisingly, after spending some time on the computer, Alex came back with a smile on his face: "Dr. Pearson, it wasn't a total waste after all. No matter how you slice it, school B is a clear winner over school A; barring sampling errors, it gives students a 50% greater chance of passing the state exam".

"What do you mean 'winner'? You talk like one of those computer freaks. In our profession, we do not talk about winners or losers, we talk about the data and let our clients jump to their own conclusions. Let's stick to the data, Alex".

"I am talking about the data", Alex replied. "What I mean is that if all the subjects we selected had remained in their assigned schools, the percentage of success in group B (school B) would have been exactly 50% higher than that of group A (school A)".

"You are a hopeless case, Alex. What's all this talk about 'had remained' and 'would have been'? Where the hell did you take your statistics 1 class? Next you are going to pretend the data tells you how many would pass the exam if we were to convince all our subjects to enroll in school B, right?"

"50%", said Alex.

"Here you are, a perfect oracle. Now how about if they enrolled in school A?" asked Dr. Pearson, somewhat amused.

"Hmm, they would all fail", said Alex, as he scanned the computer printout.

"This is too much. I give up", said Dr. Pearson. "All this talk about 'if we were' and 'if they would' is getting on my nerves. I am getting out of this place".

5.2 The Story Unfolds

As strange as it may sound, Dr. Pearson's desperation was premature, and Alex's confidence well justified. While Alex's numerical predictions were based on observing the history of those 50 students who passed the exam (i.e., on how many of them switched schools), his confidence that such precise predictions should be feasible is grounded on the facts that compliance was the same in both groups and that all subjects in one treatment arm (school A) performed uniformly (i.e., failed). The specific observation that gave rise to Alex's predictions was that none of the students who passed the exam had switched schools.

Under such observation, the data read:

$$\begin{aligned} P(y_0, d_0|z_0) = P(y_0, d_1|z_0) = P(y_0, d_0|z_1) = P(y_1, d_1|z_1) = 0.50 \\ P(y_0, d_1|z_0) = P(y_1, d_1|z_0) = P(y_0, d_1|z_1) = P(y_1, d_0|z_1) = 0.00 \end{aligned} \tag{10}$$

where z_1, d_1, and y_1 stand for "B-assigned", "B-attended", and "exam passed", respectively, and z_0, d_0, and y_0 stand for their complements. Substituting these figures in Eqs. (6) and (7) yields equality between $U(d)$ and $L(d)$, with

$$\begin{aligned} E[P(y_1|d_0, u)] &= 0.00 \\ E[P(y_1|d_1, u] &= 0.50 \\ \alpha &= 0.50 \end{aligned}$$

Thus, despite a huge rate of 50% noncompliance, the effectiveness of both programs can be determined precisely, fully supporting Alex's predictions.

The next subsection will provide an intuitive explanation for the reasoning behind this prediction.

5.3 The Story Explained

The data consists of four groups of students denoted (A, A), (A, B), (B, A), and (B, B), where the first member of each pair denotes the school assigned and the second, the school actually attended. Assume that the data show all four groups to be of equal size and that all students in the (B, B) group and no students in the other groups passed the exam.

That this data supports Alex's claims can be seen from the following:

1. First we will show that those who switched from A to B would switch from B to A had they been assigned to B. In principle, the (A, B) group may consist of two types of students: those who would stay in B if assigned to B, and those who would switch over to A if assigned to B. Call the former *contrarians* (always switching) and the latter *B-bound* (selecting B regardless of assignment). We argue that, barring sampling errors, there can be no B-bound students in (A, B), only contrarians. Indeed, if there were a fraction p of B-bound students in the population, then, due to randomization, the same fraction p is expected to exist in the two assignment arms. However, all B-bound students assigned to B must (by definition) belong to the (B, B) group, whose members all passed the exam. In contrast, all B-bound students in the (A, B) group failed the exam despite having been exposed to the same education (school B) as the B-bound students from the (B, B) group. This disparity in performance can only be reconciled by assuming that $p = 0$, namely, that there are no B-bound students at all and that all members of (A, B) are contrarians.

2. Next we will show that, barring sampling errors, members of (A, B) would all fail the exam had they stayed in A. Indeed, if all members of (A, B) are contrarians, then there are 50% contrarians among those assigned to school A, and, by randomization, the same percentage of contrarians must exist among those assigned to school B. Thus, all members of (B, A) must be contrarians as well, not A-bound, of precisely the same stock as members of (A, B). Now, to find out how effective school A is for contrarians, we observe that all members of (B, A) failed in school A. We conclude, therefore, that members of (A, B) will fail as well, being no different than members of (B, A). This substantiates Alex's first claim that no change of performance is expected by persuading all students to remain in their assigned schools.

3. It remains to be shown that students in (A, A) are identical to those in (B, B) and, hence, would pass the exam had they been assigned to B. Here we invoke the same arguments as given before to show that (A, A) could not consist of any A-bound students, that is, (A, A) students would stay in B had they been assigned B. If there were any A-bound students in (A, A) they must also be present in (B, A), but we have shown that the entire (B, A) group is made up of contrarians, which yields a contradiction. Thus, we conclude that the entire

(A, A) group must consist of pure compliers, and should perform the same as (B, B) if assigned to school B. This complete our explanation of Dr. Pearson's episode.[4]

The reader may appreciate that the reasoning behind our explanation, although logically valid, is not simple; we would not expect Dr. Pearson to do this kind of reasoning in his head on a daily basis. It is to save such labor that the human race has invented mathematical analysis. Fortunately, clinicians can now replace mental exercises with one simple formula, as given in Eq. (6).

6 Can the Model be Validated?

It is well known that one cannot infer causation from nonrandomized studies unless one is willing to make some causal assumptions. This paper appears to be claiming that causal effects can be inferred (or at least bounded) from nonrandomized studies, which raises the following two questions. What are the causal assumptions that have entered into our analysis? Can these assumptions be validated experimentally?

As defined in Section 2, our model rests on two assumptions: (a) Z is randomized, and (b) Z has no direct effect on Y. These two assumptions are equivalent to stating that Z is independent of U, a condition that economists call *exogeneity*, and which qualifies Z as an *instrumental variable* [4] relative to the relation between D and Y.[5] For a long time, it has been thought that it is impossible to verify experimentally whether a variable Z is exogenous, since the definition involves unobservable factors (or disturbances, as they are usually called) such as those represented by U. The notion of exogeneity, like that of causation itself, has been viewed as a product of subjective modeling judgment and not as an objective property that can be tested by experimental methods.

The bounds derived in this paper tell a different story: despite its elusive nature, exogeneity can be given an empirical test. The test is not guaranteed to detect all violations of exogeneity but it can, in certain circumstances, screen out very bad would-be instruments.

The empirical test dictated by our analysis can be obtained from Eq. (6) (or Eq. (18) of the Appendix). By insisting that each upper bound be higher than the corresponding lower bound, we obtain the following testable constraints on the

[4]We leave it to the reader to show that precise determination of treatment effectiveness is feasible whenever (a) the percentage of subjects switching from A to B is the same as those switching from B to A and (b) the success rate in one treatment arm is either zero or 100%.

[5]Instrumental variables is a technique invented by the geneticist Sewal Wright (1928) to help economists identify elasticities of supply and demand [8]. The key idea is that the coefficient b in the (causal) equation $Y = bX + U$ cannot be identified if X and U are correlated. However, if a third variable Z that is correlated with X but not with U can be identified, then b can be determined from the correlations between Z, X, and Y, which yields $b = R_{yz}/R_{xz}$.

observed distribution:

$$\begin{aligned}
P(y_0, d_0|z_0) + P(y_1, d_0|z_1) &\leq 1 \\
P(y_0, d_1|z_0) + P(y_1, d_1|z_1) &\leq 1 \\
P(y_1, d_0|z_0) + P(y_0, d_0|z_1) &\leq 1 \\
P(y_1, d_1|z_0) + P(y_0, d_1|z_1) &\leq 1
\end{aligned} \tag{11}$$

If any of these inequalities is violated, the investigator can safely deduce that at least one of the assumptions underlying our model is violated as well. If the assignment is carefully randomized, then any violation of these inequalities must be attributed to some direct influence that the assignment process has on subjects' responses (e.g., a traumatic experience). Alternatively, if one can eliminate direct effects of Z on Y, say through an effective use of placebo, then any observed violation of the inequalities can safely be attributed to spurious correlation between Z and U, namely, to selection bias.

The inequalities in Eq. (11), when generalized to multivalued variables, assume the form

$$\max_d \sum_y [\max_z P(y, d|z)] \leq 1 \tag{12}$$

which was called the *instrumental inequality* in [15]. We see that the instrumental inequality is violated when the controlling instrument Z manages to produce significant changes in the response variable Y while the treatment D remains constant. Although such changes could in principle be explained by strong correlations between U, D, and Y (since D does not screen off Z from Y), the instrumental inequality sets a limit on the magnitude of the changes.

The similarity to Bell's inequality in quantum physics [6, 20] is not accidental; both inequalities delineate a class of observed correlations that cannot be explained by hypothesizing latent common causes. The instrumental inequality can, in a sense, be viewed as a generalization of Bell's inequality for cases where direct causal connection is permitted to operate between the correlated observables D and Y.

The instrumental inequality can be tightened appreciably if we are willing to make additional assumptions about subjects' behavior, for example, that no individual can be discouraged by the encouragement instrument, or, mathematically, that for all u we have

$$P(d_1|z_1, u) \geq P(d_1|z_0, u)$$

Such an assumption amounts to having no contrarians in the population, namely, no subjects who will act in a spiteful manner, contrary to their assignment. Under this assumption, the inequalities in Eq. (11) can be tightened [2] to give

$$\begin{aligned}
P(y, d_1|z_1) &\geq P(y, d_0|z_0) \\
P(y, d_0|z_0) &\geq P(y, d_1|z_0)
\end{aligned} \tag{13}$$

for all $y \in \{y_0, y_1\}$. Violation of these inequalities now means either selection bias or direct effect of Z on Y or the presence of defiant subjects.

7 Concluding Remarks

Intimidated by philosophical debates, undisciplined practice, and inadequate notation, many statisticians have been reluctant to tackle problems involving causal considerations. As a consequence, statistical problems connected with treatment evaluation, liability determination, and policy decisions have often been left to antiquated methods and ad hoc analysis. Intent-to-treat analysis of noncompliance in clinical trials is one example of such antiquated methods, but, unfortunately, it continues to dominate current practices of treatment evaluation.

The results reported in this paper, a fallout of artificial intelligence research in causal reasoning, supplement existing methodologies in several ways. First, the bounds provided should allow traditional intent-to-treat analysts to evaluate how far $\Delta(Y|Z)$ is from the actual treatment effect and whether efforts to enforce compliance are likely to decrease or increase the overall treatment effectiveness. Second, thousands of extant clinical databases that were collected at great expense but either, as in the case of Dr. Pearson, abandoned or improperly analyzed can now be filtered through the bounds of Eqs. (5) to (7), which should pay off in valuable clinical knowledge. Last, and not the least promising, the availability of these bounds should encourage the deliberate design of indirect experimentation where randomized controlled experiments are infeasible or undesirable.

Another set of results of possible interest to experimental researchers are those concerning the deduction of causal effects from purely observational studies. Given an arbitrary causal graph of the type described in Figure 1, only some of whose nodes are observable, it is now possible to determine by graphical techniques whether the causal effect of one variable on another can be computed from nonexperimental data over the observables [16, 17]. If the answer is yes, then randomized experiments are not necessary and one can predict the effect of interventions by symbolic manipulation of graphs and probabilities. If the answer is no, the method may suggest surrogate experiments that are either less objectionable or more economical than brute-force randomization of the putative cause. Alternatively, bounds similar to the ones discussed in this paper can be derived.

This graphical method has uncovered many new structures that permit the identification of causal effects from nonexperimental observations. For example, the structure of Figure 2 represents a class of observational studies in which the causal effect of X on Y can be determined by measuring a variable Z that *mediates* the interaction between the cause and its effect. This stands contrary to most of the literature on statistical experimentation, which considers the measurement of intermediate variables affected by the action to be useless, if not harmful, for causal inference [5, 18]. The relevance of such structures in practical situations can be seen, for instance, if we identify X with smoking, Y with lung cancer, Z with the amount of tar deposited in a subject's lungs, and U with an unobserved carcinogenic genotype that, according to the tobacco industry, also induces an inborn craving for nicotine. In this case, it is possible to determine graphically that measurement of Z renders randomization unnecessary, which provides us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer. (Assuming, of course, that the data $P(x,y,z)$ is made available and that we believe that smoking does not have any direct causal effect on lung cancer except that mediated by tar deposits.) Moreover, the causal effect of X on Y can be written in close mathematical form as

$$E[P(y_1|x,u)] = \sum_z P(z|x) \sum_{x'} P(y|x',z')P(x')$$

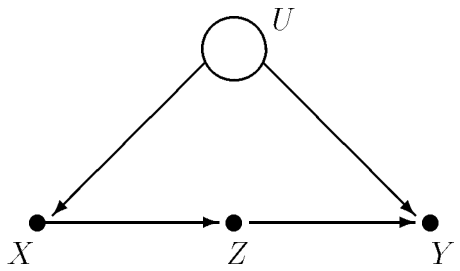

Figure 2: *Causal inference using an intermediate variable Z.*

Acknowledgment

This investigation owes much to the results of Alex Balke, as reported in [2, 3]; Alex's encounter with Dr. Pearson was not entirely fantastical. I value the encouragement of James Robins, who probably is the only public health researcher who currently appreciates the usefulness of these results. I partially thank colleagues, reviewers, and associate editors who have insisted that "intent-to-treat is the only legitimate way of analyzing noncompliance data" and that "the proposed method undermines the importance of randomization". Were it not for such objections, I would not have found out how crude traditional methods are. This research was partially supported by Air Force grant #F49620-94-1-0173, NSF grant #IRI-9200918, and Northrop-Rockwell Micro grant #93-124.

References

[1] J.D. Angrist and G.W. Imbens, Source of identifying information in evaluation models, Discussion Paper 1568, Department of Economics, Harvard University, Cambridge, MA, 1991.

[2] A.A. Balke and J. Pearl, Nonparametric bounds on causal effects from partial compliance data, Technical Report No. 199, Cognitive Systems Laboratory,

UCLA Computer Science Department, Los Angeles, CA, September 1993. Submitted for publication.

[3] A.A. Balke and J. Pearl, Counterfactual probabilities: Computational methods, bounds, and applications, in: R. Lopez de Mantaras and D. Poole, eds., *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Mateo, CA, 1994) 46-54.

[4] R.J. Bowden and D.A. Turkington, *Instrumental Variables* (Cambridge University Press, Cambridge, UK, 1984).

[5] D.R. Cox, *The Planning of Experiments* (John Wiley and Sons, New York, 1958).

[6] J.T. Cushing and E. McMullin, eds., *Philosophical Consequences of Quantum Theory* (University of Notre Dame Press, South Bend, IA, 1989).

[7] B. Efron and D. Feldman, Compliance as an explanatory variable in clinical trials, *Journal of the American Statistical Association*, Vol. 86, No. 413 (1991) 9–26.

[8] A.S. Goldberger, Structural equation methods in the social sciences, *Econometrica*, Vol. 40 (1972) 979–1001.

[9] J.J. Heckman, Randomization and social policy evaluation, in: C. Manski and I. Garfinkle, eds., *Evaluations of Welfare and Training Programs* (Harvard University Press, 1992) 201–230.

[10] M.S. Kramer and S. Shapiro, Scientific Challenges in the Application of Randomized Trials, *Journal of the American Medical Association* Vol. 252 (1984) 2739–2745.

[11] The Lipid Research Clinics Coronary Primary Prevention Trial results, parts I and II, *Journal of the American Medical Association*, Vol. 251, No. 3 (1984) 351–374.

[12] C.F. Manski, Nonparametric bounds on treatment effects, *American Economic Review, Papers and Proceedings*, Vol. 80 (1990) 319–323.

[13] J. Palca, AIDS drug trials enter new age, *Science Magazine* (1989) 19–21.

[14] J. Pearl, From Bayesian networks to causal networks, in: *Proceedings of the Adaptive Computing and Information Processing Seminar,* Brunel Conference Centre, London, January 25-27, 1994. See also *Statistical Science*, Vol. 8, No. 3 (1993) 266–269.

[15] J. Pearl, A note on testing exogeneity of instrumental variables, Technical Report R-211-S, Cognitive Systems Laboratory, UCLA Computer Science Department, Los Angeles, CA, December 1993.

[16] J. Pearl, Causal diagrams for experimental research, Technical Report R-218-L, Cognitive Systems Laboratory, UCLA Computer Science Department, Los Angeles, CA, May 1993. To appear in *Biometrika*.

[17] J. Pearl, A probabilistic calculus of actions, in: R. Lopez de Mantaras and D. Poole, eds., *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufman, San Mateo, CA, 1994) 454–462.

[18] J.W. Pratt and R. Schlaifer, On the interpretation and observation of laws, *Journal of Econometrics*, Vol. 39 (1988) 23–52.

[19] J.M. Robins, The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in: L. Sechrest, H. Freeman, and A. Mulley, eds., *Health Service Research Methodology: A Focus on AIDS* (NCHSR, U.S. Public Health Service, 1989) 113–159.

[20] P. Suppes, Probabilistic causality in space and time, in: B. Skyrms and W.L. Harper, eds., *Causation, Chance, and Credence* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988) 135–151.

[21] S. Wright, Appendix, in: P.G. Wright, *The Tariff on Animal and Vegetable Oils* (Macmillan, New York, 1928) PAGES.

Appendix

To prove Eq. (5), we write

$$P(y,d|z) = \sum_u P(y|d,u)\ P(d|z,u)\ P(u) \tag{14}$$

and define the following four functions:

$$f_0(u) = P(y_1|d_0,u) \qquad g_0(u) = P(d_1|u,z_0) \tag{15}$$

$$f_1(u) = P(y_1|d_1,u) \qquad g_1(u) = P(d_1|u,z_1) \tag{16}$$

This permits us to express six independent components of $P(y,d|z)$ as expectations of these functions:

$$\begin{array}{rcl} P(y_1,d_0|z_0) & = & E[f_0(1-g_0)] = a \\ P(y_1,d_0|z_1) & = & E[f_0(1-g_1)] = b \\ P(d_1|z_0) & = & E(g_0) = c \\ P(d_1|z_1) & = & E(g_1) = d \\ P(y_1,d_1|z_0) & = & E[f_1 \cdot g_0] = e \\ P(y_1,d_1|z_1) & = & E[f_1 \cdot g_1] = h \end{array} \tag{17}$$

For any two random variables X and Y such that $0 \leq X \leq 1, 0 \leq Y \leq 1$ we have

$$1 + E(XY) - E(Y) \geq E(X) \geq E(XY)$$

since $E[(1-X)(1-Y)] \geq 0$. This inequality holds for any pair of f, g functions (since they lie between 0 and 1) and we can write:

$$\begin{array}{c} 1 + E(f_1 g_0) - E(g_0) \geq E(f_1) \geq E(f_1 g_0) \\ 1 + E(f_1 g_1) - E(g_1) \geq E(f_1) \geq E(f_1 g_1) \\ 1 + E[f_0(1-g_0)] - E(1-g_0) \geq E(f_0) \geq E[f_0(1-g_0)] \\ 1 + E[f_0(1-g_1)] - E(1-g_1) \geq E(f_0) \geq E[f_0(1-g_1)] \end{array}$$

or

$$\begin{array}{l} \max[h;e] \leq E(f_1) \leq \min[(1+e-c);(1+h-d)] \\ \max[a;b] \leq E(f_0) \leq \min[(a+c);(b+d)] \end{array} \tag{18}$$

Lower bounding $E(f_1)$ and upper bounding $E(f_0)$ provides a lower bound for their difference:

$$\begin{aligned} E(f_1) - E(f_0) &\geq \max[e;h] - \min[(a+c);(b+d)] \\ &\geq h - (a+c) \end{aligned} \tag{19}$$

Substituting back the $P(y,d|z)$ expressions from Eqs. (15) to (17) yields the lower bound of Eq. (5). Similarly, the difference can be upper bounded by

$$\begin{aligned} E(f_1) - E(f_0) &\leq \min[(1+e-c);(1+h-d)] - \max[a;b] \\ &\leq 1 + h - d - a \end{aligned}$$

thus proving Eq. (5).

To evaluate

$$\alpha^* \;=\; E\{[P(y_1|d_1,u) - P(y_1|d_0,u)]|D \;=\; d_1\}$$

we define

$$\begin{aligned}\Delta(u) \;&\equiv\; P(y_1|d_1,u) - P(y_1|d_0,u) \;=\; f_1(u) - f_0(u)\\ q \;&\equiv\; P(z_1)\end{aligned}$$

and write

$$\begin{aligned}\alpha^* &= E[\Delta(u)|D = d_1]\\ &= \textstyle\sum_u \Delta(u)P(u|d_1)\\ &= \tfrac{1}{P(d_1)}\textstyle\sum_u \Delta(u)P(d_1|u)P(u)\\ &= \tfrac{1}{P(d_1)}\textstyle\sum_u\sum_z \Delta(u)P(d_1|u,z)P(z)P(u)\\ &= \tfrac{1}{P(d_1)}\textstyle\sum_u \Delta(u)P(u)[P(z_1)g_1(u) + P(z_0)g_0(u)]\\ &= \tfrac{1}{P(d_1)}E\{[f_1(u) - f_0(u)][qg_1(u) + (1-q)g_0(u)]\}\\ &= \tfrac{1}{P(d_1)}E[qf_1g_1 + (1-q)f_1g_0 - qf_0g_1 - (1-q)f_0g_0]\\ &= \tfrac{1}{P(d_1)}[qh + (1-q)e - qE(f_0g_1) - (1-q)E(f_0g_0)]\\ &= \tfrac{1}{P(d_1)}[qh + (1-q)e - q(E(f_0) - b) - (1-q)(E(f_0) - a)]\\ &= \tfrac{1}{P(d_1)}[q(h+b) + (1-q)(e+a) - E(f_0)]\end{aligned} \tag{20}$$

Substituting the expressions for $(h+b)$ and $(e+a)$ from Eq. (17) and using

$$a \leq E(f_0) < a + c$$

from Eq. (18), we obtain upper and lower bounds on α^*:

$$\begin{aligned}\frac{1}{P(d_1)}[P(y_1) - P(d_1|z_0) - P(y_1,d_0|z_0)] \leq \alpha^*\\ \frac{1}{P(d_1)}[P(y_1) - P(y_1,d_0|z_0)] \geq \alpha^*\end{aligned} \tag{21}$$

Alternatively, collecting common terms in both expressions of Eq. (21), we get

$$-\frac{P(y_0,d_1|z_0)}{P(d_1)} \;\leq\; \alpha^* - \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1)/P(z_1)} \;\leq\; \frac{P(y_1,d_1|z_0)}{P(d_1)}$$

Thus,

$$\alpha^* = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1)}$$

if $P(d_1|z_0) = 0$, which proves Eqs. (8) and (9).