TECHNICAL REPORT
R-431-L
October 2014

Causes of Effects and Effects of Causes*

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

October 6, 2014

Abstract

This paper summarizes methods that were found useful in estimating the probability that one event was a necessary cause of another, as interpreted by law makers. We show that the fusion of observational and experimental data can yield informative bounds which, under certain circumstances, meet legal criteria of causation. We further investigate the circumstances under which such bounds can emerge, and the philosophical dilemma associated with determining individual cases from statistical data.

1 Introduction

In many areas of the physical and social sciences interest focuses on identifying causes of effect (CoE) rather than predicting effects of causes (EoC). This distinction assumes critical importance in legal settings where individual cases are to be decided, and population behavior is of secondary importance. The fact that standard statistics deals with inference from population data, not with individual cases has led some to speculate that the problem of deciding

*This paper was motivated by anachronistic claims made by Dawid et al. (2014a) which were brought to my attention by Nicholas Jewell. I subsequently wrote two rebuttals (Pearl, 2014a) and (Pearl, 2014b) which were based on chapter 9 of my book (Pearl, 2000a) and additional analyses presented in sections 5 and 6 of this paper. This paper is a self-contained summary of what I consider to be the state of the art in this important problem of Causes of Effect.

causes of effects may reside beyond the realm of statistical inference, perhaps even beyond the province of the empirical sciences (Dawid, 2000; Shafer, 2000; Rubin, 2014). A recent article by Dawid, Faigman, and Fienberg (2014a) reiterates those speculations, and attempts to revive old doubts as to the ability of principled scientific methods to ever provide answers to CoE questions, such as those raised in legal settings.

My aim in this paper is to share with readers a progress report on what has been accomplished on the question of "causes of effects," CoE, how far we have come in using population data to decide individual cases, and how well we can answer questions that law makers ask about individual's guilt or innocence. I hope this account convinces readers that the analysis of "causes of effects," CoE, has not lagged behind that of EoC. Both modes of reasoning enjoy a solid mathematical basis, endowed with powerful tools of analysis, and researchers on both fronts now possess solid understanding of applications, identification conditions, and estimation techniques.

I begin this account with a brief exposition of counterfactuals, what they stand for, how they are computed, how they are assigned probabilities and how they are estimated from a scientific model of reality (Section 2). Next (Section 3), I discuss a simple example of a law suit in which a legal requirement demands the estimation the probability of the counterfactual sentence: "Mr. A would have been alive had he not taken this drug," and I cite a few theoretical results that permit us to bound this "probability of necessity" from a combination of experimental and observational data. Section 4, illustrates the use of these bounds on fictitious data and shows that, under certain circumstances, data may dictate unexpected conclusions, for instance, that the defendant is guilty "with probability one." Motivated by such extreme cases, Section 5 uncovers general laws that govern necessary causation and how it is informed by empirical findings. In particular, we show that, regardless of confounding, the gap between the upper and lower bounds is estimable from one statistical parameter – the ratio of non-responses to responses in similar situations. In Section 6 we describe how bounds on the probability of necessity emerge from a specific data-generating process, and what model parameters affect the resulting bounds. Finally, Section 7 discusses the legal and cognitive question of whether juries can/should be persuaded to heed to statistical evidence in single case settings, how they should interpret such evidence, and whether arguments for such extreme findings as "guilty with probability one" could ever be convincing.

2 The Logic of Counterfactuals

A good place to start is the mathematization of counterfactuals, a development that is responsible, at least partially, for legitimizing counterfactuals in scientific discourse,[1] and which has reduced the quest for "causes of effects" to an exercise in logic.

At the center of this logic lies a model, M, consisting of a set of equations similar to those used by physicists, geneticists (Wright, 1921) economists (Haavelmo, 1943) and social scientists (Duncan, 1975) to articulate scientific knowledge in their respective domains. M consists of two sets of variables, U and V, and a set F of equations that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u, of all variables in V and U and, accordingly, *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables corresponds to defining a "unit," or a "situation" in the model, and uniquely determines the values of all variables in V. Therefore, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V. The probabilities on U and V can best be interpreted as the proportion of the population with a particular combination of values on U and/or V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$, where Y and X are any variables in V. The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x" as an instruction to make a minimal modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x, which may be thought of as an external action $do(X = x)$, not necessarily by a human experimenter, that imposes the condition $X = x$. This replacement permits the constant x to differ from the actual value of X (namely $f_x(v, u)$) without rendering the system of equations inconsistent, thus allowing all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M, with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads

[1] DFF's article makes generous use of counterfactuals, which attests to the impact of this development. For discussions concerning the place of counterfactuals in science, including their role in defining "causes of effects" see (Dawid, 2000; Pearl, 2000b).

(Balke and Pearl, 1994a,b):

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{1}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models. (see also Pearl, 2009, Chapter 7). They showed that the axioms governing recursive structural counterfactuals are identical to those used in the potential outcomes framework, hence the two systems are logically identical – a theorem in one is a theorem in the other. This means that relying on structural models as a basis for counterfactuals does not impose additional assumptions beyond those routinely invoked by potential outcome practitioners. Consequently, going from effects to causes does not require extra mathematical machinery beyond that used in going from causes to effects.

Since our model M consists of a set of structural equations, it is possible to calculate probabilities that might at first appear nonsensical. As noted above the probability distribution on U, $P(u)$, induces a well defined probability distribution on V, $P(v)$. As such, it not only defines the probability of any single counterfactual, $Y_x = y$, but also the joint distribution of all counterfactuals. As also noted above these probabilities refer to the proportion of individuals in the population with specific *counterfactual* values that may or may not be observed. Thus the probability of the Boolean combination, "$Y_x = y$ AND $Z_{x'} = z$" for variables Y and Z in V and two different values of X, x and x', is well-defined even though it is impossible for both outcomes to be simultaneously observed as $X = x$ and $X = x'$ cannot be concurrently true.

To answer CoE type questions, such as "if X were x_1 would Y be y_1 for individuals for whom in fact X is x_0 and Y is y_o" we need to compute the conditional probability $P(Y_{x_1} = y_1|Y = y_0, X = x_0)$. This probability, that is the proportion of the population with this combination of counterfactual values, is well-defined once the structural equations and the distribution of exogenous variables, U, is known.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any information about an individual, can be computed by the 3-step process:

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

Pearl (2000a, pp. 296–299; 2012) gives several examples illustrating the simplicity of this computation and how CoE-type questions can be answered when the model M is known. If M is not known, but is assumed to take a parametric form, one can use population data to estimate the parameters and, subsequently, all counterfactual queries can be answered, including those that pertain to causes of individual cases (Pearl, 2009, pp. 389–391; 2012). Thus the challenge of reasoning from group data to individual cases has been met.

When the model M is not known, we can prove that, in general, probabilities of causes are not identifiable from experimental, or observational data. However, by combining experimental and observational group data with observations about an individual, tight bounds can be derived, which can be quite informative, often satisfying legal criteria for CoE.

We will illustrate these bounds it an example taken from judicial context similar to the one considered by DFF.

3 Legal Responsibility from Experimental and Nonexperimental Data

A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve back pains. The manufacturer claims that experimental data on patients with back pains show conclusively that drug x may have only minor effect on death rates. However, the plaintiff argues that the experimental study is of little relevance to this case because it represents average effects on *all* patients in the study, not on patients like Mr. A who did not participate in the study. In particular, argues the plaintiff, Mr. A is unique in that he used the drug on his own volition, unlike subjects in the experimental study who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes nonexperimental data on patients who, like Mr. A, chose drug x to relieve back pains, but were not part of any experiment. The court must now decide, based on both the experimental and nonexperimental studies, whether it is "more probable than not" that drug x was in fact the cause of Mr. A's death.

This example falls under the category of "causes of effects" because it concerns situation in which we observe both the effect, $Y = y$, and the putative cause $X = x$ and we are asked to assess, counterfactually, whether the former would have occurred absent the latter.

Assuming binary events, with $X = x$ and $Y = y$ representing treatment and outcome, respectively, and $X = x'$, $Y = y'$ their negations, our target quantity can be formulated directly from the English sentence:

> "Find the probability that if X had been x', Y would be y', given that, in reality, X is x and Y is y."

to give:

$$PN(x, y) = P(Y_{x'} = y' | X = x, Y = y) \tag{2}$$

This counterfactual quantity, which Robins and Greenland (1989) named "probability of causation" and Pearl (2000a, p. 296) named "probability of necessity" (PN), to be distinguished from two other nuances of "causation," captures the "but for" criterion according to which judgment in favor of a plaintiff should be made if and only if it is "more probable than not" that the damage would not have occurred *but for* the defendant's action (Robertson, 1997). In contrast, the "probability of causation" (PC) measure proposed by Dawid, Fienberg, and Faigman:

$$PC = P(Y_{x'} = y' | Y_x = y)$$

represents the probability that a person who took the drug under experimental conditions and died, $Y_x = y$, would be alive had he not been assigned the drug, $Y_{x'} = y'$. It thus represents the probability that the drug was the cause of death of a subject who died in the experimental setup. Very few court cases deal with deaths under experimental circumstances; most deal with deaths, damage, or injuries that took place under natural, every day conditions, for which the DFF's measure is inapplicable.[2]

Having written a formal expression for PN, Eq. (2), we can move on to the identification phase and ask what assumptions would permit us to identify PN from empirical studies, be they observational, experimental or a combination thereof.

This problem was analyzed in Pearl (2000a, Chapter 9) and yielded the following results:

[2]For additional discussions of DFF's proposal, see footnote 4, Appendix A, and Pearl (2014b).

Theorem 1 *If Y is monotonic relative to X, i.e., $Y_1(u) \geq Y_0(u)$, then* PN *is identifiable whenever the causal effect $P(y|do(x))$ is identifiable and, moreover,*

$$PN = \frac{P(y) - P(y|do(x'))}{P(x,y)} \tag{3}$$

or,[3]

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}. \tag{4}$$

The first term on the r.h.s. of (4) is the familiar excess risk ratio (ERR) that epidemiologists have been using as a surrogate for PN in court cases (Cole, 1997; Greenland, 1999; Robins and Greenland, 1989). The second term represents a *correction* needed to account for confounding bias, that is, $P(y|do(x')) \neq P(y|x')$ or, put in words, when the proportion of population for whom $Y = y$ when X is set to x' for everyone is not the same as the proportion of the population for whom $Y = y$ among those observed to acquire the value $X = x'$.

Equation (4) thus provides a more refined measure of causation, which can be used for monotonic $Y_x(u)$ whenever the causal effect $P(y|do(x))$ can be estimated, from either randomized trials or graph-assisted observational studies (e.g., through the back-door criterion (Pearl, 1993) or the *do*-calculus). More significantly, it has also been shown (Tian and Pearl, 2000) that the expression in (3) provides a lower bound for PN in the general, nonmonotonic case. In particular, the tight upper and lower bounds on PN are given by:

$$\max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq \min\left\{1, \frac{P(y'|do(x')) - P(x',y')}{P(x,y)}\right\} \tag{5}$$

In drug-related litigation, it is not uncommon to obtain data from both experimental and observational studies. The former is usually available at the manufacturer or the agency that approved the drug for distribution (e.g., FDA), while the latter is easy to obtain by random surveys of the population. If it is the case that the experimental and survey data have been drawn at random from the same population, then the experimental data can be used to estimate the counterfactuals of interest, e.g., $P(Y_x = y)$ for the observational as well as experimental sampled populations. In such cases, the standard lower bound used by epidemiologists to establish legal responsibility, the Excess Risk Ratio,

[3]Equation (4) is obtained from (3) by writing $P(y) = P(y|x)P(x) + P(y|x')(1 - P(x))$.

can be improved substantially using the corrective term of Eq. (4). Likewise, the upper bound of Eq. (5) can be used to exonerate drug-makers from legal responsibility. Cai and Kuroki (2006) analyzed the finite-sample properties of PN. Yamamoto (2012) used instrumental variables to derive similar bounds for subpopulations permitting effect identification.

4 Numerical Example

To illustrate the usefulness of the bounds in Eq. (5), consider the (hypothetical) data associated with the two studies shown in Table 1. (In the analyses below, we ignore sampling variability, that is, we assume that our population is of infinite size.)

	Experimental		Nonexperimental	
	$do(x)$	$do(x')$	x	x'
Deaths (y)	16	14	2	28
Survivals (y')	984	986	998	972

Table 1:

The experimental data provide the estimates

$$P(y|do(x)) = 16/1000 = 0.016, \tag{6}$$
$$P(y|do(x')) = 14/1000 = 0.014; \tag{7}$$

while the nonexperimental data provide the estimates

$$P(y) = 30/2000 = 0.015, \tag{8}$$
$$P(y,x) = 2/2000 = 0.001, \tag{9}$$
$$P(y|x) = 2/1000 = 0.002, \tag{10}$$
$$P(y|x') = 28/1000 = 0.028. \tag{11}$$

Assuming that drug x can only cause (but never prevent) death, monotonicity holds and Theorem 1 (Eq. 4) yields

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)} =$$
$$= \frac{0.002 - 0.028}{P(y|x)} + \frac{0.028 - 0.014}{0.001} = -13 + 14 = 1 \tag{12}$$

We see that while the observational excess risk ratio ERR is negative (-13), giving the impression that the drug is actually preventing deaths, the bias-correction term ($+14$) rectifies this impression and sets the probability of necessity (PN) to unity. Moreover, since the lower bound of Eq. (5) becomes 1, we conclude that $PN = 1.00$ even without assuming monotonicity. Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug x was in fact responsible for the death of Mr. A. Note that DFF's proposal of using the experimental excess risk ratio would yield a much lower result:

$$\frac{P(y|do(x)) - P(y|do(x'))}{P(y|do(x))} = \frac{0.016 - 0.014}{0.016} = 0.125. \tag{13}$$

which does not meet the "more probable than not" requirement.[4]

What the experimental study does not reveal is that, given a choice, terminal patients tend to avoid drug x. That is, the 14 patients in the experimental study who did not take the drug and died anyway would have avoided the drug if they were in the nonexperimental study. In fact, as our analysis above shows, there are no terminal patients who would choose x (given the choice). If there were terminal patients that would choose x, given the choice, then by randomization some of these patients (50% in our example) would be in the control group in the experimental data. As a result, the proportion of deaths in the control group in the experimental data, $P(y_{x'})$ would be higher than the proportion of terminal patients in the nonexperimental data, $P(y, x')$. However, examining the data in our hypothetical example, we observe that $P(y_{x'}) = P(y, x') = .0014$ implying that there are no terminal patients in the nonexperimental data who choose the treatment condition. As such, any individual in the nonexperimental data who choose the treatment and died, must have died because of the treatment as they were not terminal.

The numbers in Table 1 were obviously contrived to represent an extreme case and so facilitate a qualitative explanation of the validity of (12). Nevertheless, it illustrates decisively that a combination of experimental and nonexperimental studies may unravel what experimental studies alone will not reveal

[4]The difference between DFF's PC and PN represents not merely an improvement of bounds but a profound conceptual difference in what the correct question is for CoE. Using DFF's notation we have $PC = Pr(R_0 = 0|R_1 = 1)$ and $PN = Pr(R_0 = 0|A = 1, R = 1)$. PC is the wrong measure to use in legal context because the conditioning event $R_1 = 1$ does not imply that the action $A = 1$ was actually executed. Moreover, PC does not take into account the possibility that plaintiffs who chose the treatment voluntarily are more likely to be in need of such treatment, as well as more capable of obtaining it. The same goes for personal decision making; PC does not take into account the fact that, if I took aspirin and my headache is gone, I am the type of person who expects aspirin to help my headache. Formally, while $A = 1$ and $R = 1$ imply $R_1 = 1$ the converse does not hold; the former is the more specific reference class.

and, in addition, that such combination may provide a necessary test for the adequacy of the experimental procedures. For example, if the frequencies in Table 1 were slightly different, they could easily yield a PN value greater than unity in (12), thus violating consistency, $P(y|do(x)) \geq P(x, y)$. Such violation must be due to incompatibility of experimental and nonexperimental groups, or an improperly conducted experiment.

This last point may warrant a word of explanation, lest the reader wonder why two data sets—taken from two separate groups under different experimental conditions—should constrain one another. The explanation is that certain quantities in the two subpopulations are expected to remain invariant to all these differences, provided that the two subpopulations were sampled randomly from the population at large. These invariant quantities are simply the causal effects probabilities, $P(y|do(x'))$ and $P(y|do(x))$. Although these probabilities were not measured in the observational group, they must nevertheless be the same as those measured in the experimental group (ignoring differences due to sampling variability). The invariance of these quantities implies the inequalities of (5).

The example of Table 1 shows that combining data from experimental and observational studies which, taken separately, may indicate no causal relations between X and Y, can nevertheless bring the lower bound of Eq. (5) to unity, thus implying causation *with probability approaching one*.

Such extreme results demonstrate that a counterfactual quantity PN which at first glance appears to be hypothetical, ill-defined, untestable and, hence, unworthy of scientific analysis is nevertheless definable, testable and, in certain cases, e.g., when monotonicity holds, even identifiable. Moreover, the fact that, under certain combinations of data, and making no assumptions whatsoever, an important legal claim such as "the plaintiff would be alive had he not taken the drug" can be ascertained with probability approaching one, is a remarkable tribute to formal analysis.[5]

[5] Another counterfactual quantity that has been tamed by analysis is the Effect of Treatment on the Treated (ETT): $ETT = P(Y_{x'} = y|X = x)$. Shpitser and Pearl (2009) have shown that despite its blatant counterfactual character (e.g., "I just took an aspirin, perhaps I shouldn't have?"), ETT can be evaluated from experimental studies in many, though not all cases. It can also be evaluated from observational studies whenever a sufficient set of covariates can be measured that satisfies the back-door criterion and, more generally, in a wide class of graphs that permit the identification of conditional interventions. Numerical examples of these cases, and the philosophical question they evoke, are discussed in (Pearl, 2013).

5 How informative are the PN bounds?

To see how informative the bounds are, and how sensitive they are to variations in the experimental and observational data, let us express the bounds in Eq. (5) in terms of more familial parameters, as they apply to the unit square.

A few algebraic steps allow us to express the lower bound (LB) and upper bound (UB) as:

$$\begin{aligned} LB &= ERR + CF \\ UB &= ERR + q + CF \end{aligned} \tag{14}$$

where ERR, CF, and Q are defined as follows:

$$CF \stackrel{\Delta}{=} [P(y|x') - P(y_{x'})]/P(x,y) \tag{15}$$

$$ERR \stackrel{\Delta}{=} 1 - 1/RR = 1 - P(y|x')/P(y|x) \tag{16}$$

$$q \stackrel{\Delta}{=} P(y'|x)]/P(y|x) \tag{17}$$

Here, CF (termed "confounding factor") represents the normalized degree of confounding among the unexposed ($X = x'$), ERR is the "excess risk ratio" and q is the ratio of negative to positive outcomes among the exposed.

Figures 1(a,b) depicts these bounds as a function of ERR, and reveals three rather insightful observations. First, regardless of confounding the interval $UB - LB$ remains constant and depends on only one observable parameter, $P(y'|x)/P(y|x)$. Second, when confounding is present, the lower bound may rise to meet the $PN > \frac{1}{2}$ criterion. Lastly, the amount of rise is given by CF, which is the only estimate needed from the experimental data; the causal effect $P(y_x) - P(y_{x'})$ is not needed.

Theorem 1 further assures us that, if monotonicity can be assumed, the upper and lower bounds coincide, and the gap collapses entirely, as shown in Fig. 1(b).

We thus conclude that confounding can be a blessing; it may boost the lower bound (when $CF > 0$) or lower the upper found (when $CF < 0$), and thus assist in passing or failing the "more probable than not" criterion. Confounding is in fact the only mechanism through which the idiosyncratic behavioral of an individual can be excavated from population data, assuming of course that no other information is available about the specific individual.

How does confounding provide information about PN? To answer this question, we will now examine a specific data-generating model, and trace the way our three data parameters, ERR, CF and q are determined by the model parameters. It is important to emphasize, however, that contrary to prevailing

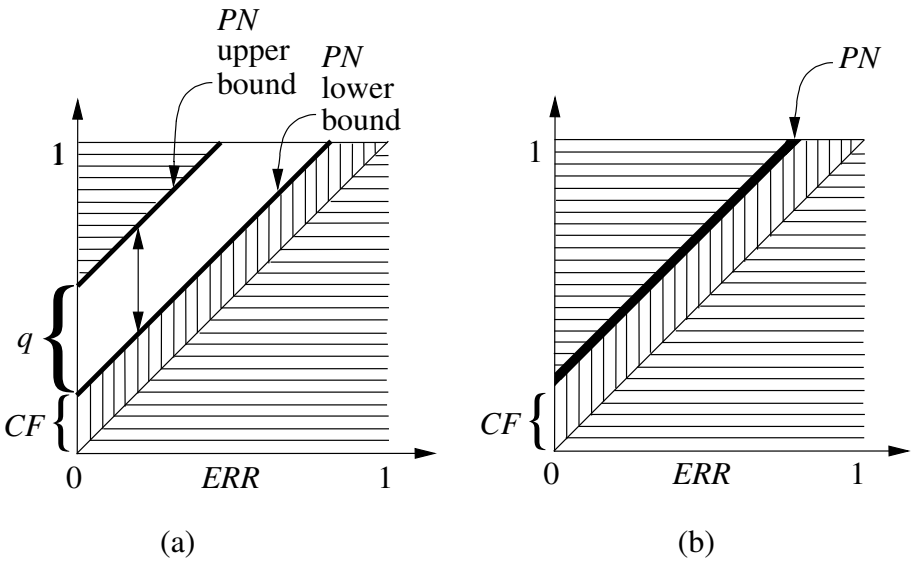

Figure 1: (a) Showing how probabilities of necessity (PN) are bounded, as a function of the Excess Risk Ratio (ERR) and the Confounding Factor (CF) (Eq. 14); (b) Showing how PN is identified when monotonicity is assumed (Theorem 1).

opinions, these bounds do not require knowledge of the data-generating model; population data from observational and experimental studies are all that is needed.

6 How the PN bounds are generated?

Consider the following example. Assume that the population of patients contains a fraction r of individuals who suffer from a certain death-causing syndrome Z, which simultaneously makes it uncomfortable for them to take the drug. Referring to Fig. 2, let $Z = z_1$ and $Z = z_0$ represent, respectively, the presence and absence of the syndrome, $Y = y_1$ and $Y = y_0$ represent death and survival, respectively and $X = x_1$ and $X = x_0$, represent taking and not taking the drug. Assume that patients carrying the syndrome, $Z = z_1$, are terminal cases, for whom death occurs with probability 1, regardless of whether they take the drug. Patients not carrying the syndrome, on the other hand, incur death with probability p_2 if they take the drug and with probability p_1 if the don't. We will further assume $p_2 > p_1$ so that the drug appears to be a risk factor for ordinary patients, and that patients having the syndrome are more likely to avoid the drug; that is, $q_2 < q_1$ where $q_1 = P(x_1|z_0)$ and $q_2 = p(x_1|z_1)$.

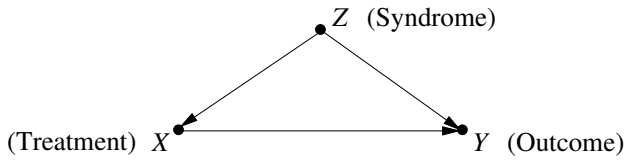

Figure 2: Model generating the experimental and observational data of Eqs. (20) and (21). Z represents an unobserved confounder affecting both treatment (X) and outcome (Y).

Based on this model, we can compute the causal effect of the drug on death using:

$$P(y|do(x)) = \sum_z P(y|x,z)P(z) \quad \text{for all } y \text{ and } x \tag{18}$$

and the joint distribution $P(x,y)$ using:

$$P(y,x) = \sum_z P(y|x,z)P(x|z)P(z) \quad \text{for all } y, x \tag{19}$$

Substituting the model's parameters and assuming $r = 1/2$ gives:

$$P(y_1|do(x)) \;\; = \begin{cases} (1+p_2)/2 & \text{for} \quad x = x_1 \\ (1+p_1)/2 & \text{for} \quad x = x_0 \end{cases} \tag{20}$$

$$P(y,x) \;\; = \begin{cases} (q_2 + p_2 q_1)/2 & \text{for} \quad x = x_1 \quad y = y_1 \\ [1 - q_2 + p_1(1-q_1)]/2 & \text{for} \quad x = x_0 \quad y = y_1 \\ (1-p_2)q_1/2 & \text{for} \quad x = x_1 \quad y = y_0 \\ (1-p_1)(1-q_1)/2 & \text{for} \quad x = x_0 \quad y = y_0 \end{cases} \tag{21}$$

Accordingly, the bounds of Eq. (5) become:

$$(p_2 - p_1)/(p_2 + q_2/q_1) \leq PN \leq (1-p_1)/(p_2 + q_2/q_1) \tag{22}$$

Equating the upper and lower bounds in (22) reveals that PN is identified if and only if $q_1(1-p_2) = 0$, namely, if patients carrying the syndrome either do not take the drug or do not survive if they do. For intermediate value of p_2 and q_1, PN is constrained to an interval that depends on all four parameters.

Figure 3 displays the lower bound (red curve) as a function of the parameter $\beta = q_2/q_1 p_2$, for $p_1 = 0$ and the upper bounds (green curves) for $p_2 = 1.00, 0.5, 0.33, 0.25$. We see that lower bound approaches 1 when q_2 approaches zero, while the upper bounds are situated a factor $1/p_2$ above the lower bound.

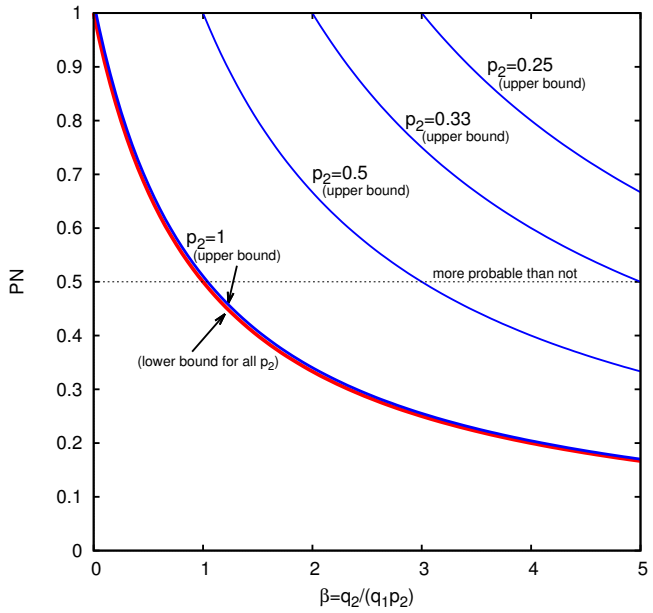

Figure 3: Showing the lower bound of PN for $p_1 = 1$ (red curve) and several upper bounds (blue curves).

It is important to emphasize at this point that the bounds shown in Fig. 3 are subsumed by the universal bounds shown in Fig. 1. In other words, regardless of the structure of the data-generating model shown in Fig. 2, and regardless of the parameters used in this model, PN is guarantees to fall between the bounds given in Eqs. (15)–(17). This means that, if one were to simulate the observed data on a computer, then, regardless of the structure of the simulator, the PN associated with that simulator will fall within the bounds shown in Fig.1.[6] If we further imagine that the simulator stands for Nature (or, say the anatomy of the injured) the argument can be interpreted to mean that there is no way for Nature to have generated the data while entailing a probability of necessity outside the LB and UP bounds. Whether juries can be persuaded by such nature-minded arguments will be discussed in the next section.

[6]Note that with every simulator defines a unique PN, though the model specified in Fig. 2 has no unique PN, because it is defined probabilistically, not functionally.

7 Is "Guilty with Probability One" ever possible?

People tend to disbelieve this possibility for two puzzling aspects of the problem:

1. that a hypothetical, generally untestable quantity can be ascertained with probability one under certain conditions;

2. that a property of an untested individual can be assigned a probability one, on the basis of data taken from sampled population.

The first puzzle is not really surprising for students of science who take seriously the benefits of logic and mathematics. Once we give a quantity formal semantics we essentially define its relation to the data, and it not inconceivable that data obtained under certain conditions would sufficiently constrain that quantity, to a point where it can be determined exactly.

The second puzzle is the one that gives most people a shock of disbelief. For a statistician, in particular, it is a rare case to be able to say anything certain about a specific individual who was not tested directly. This emanates from two factors. First, statisticians normally deal with finite samples, the variability of which rules out certainty in any claim, not merely about an individual but also about any property of the underlying distribution. This factor, however, should not enter into our discussion, for we have been assuming infinite samples throughout. (Readers should imagine that the numbers in Table 1 stand for millions.)

The second factor emanates from the fact that, even when we know a distribution precisely, we cannot assign a definite probabilistic estimate to a property of a specific individual drawn from that distribution. The reason is, so the argument goes, that we never know, let alone measure, all the anatomical and psychological variables that determine an individual's behavior, and, even if we knew, we would not be able to represent them in the crude categories provided by the distribution at hand. Thus, because of this inherent crudeness, the sentence "Mr. A would be dead" can never be assigned a probability one (or, in fact, any definite probability).

This argument, advanced by Freedman and Stark (1999) is incompatible with the way probability statements are used in ordinary discourse, for it implies that every probability statement about an individual must be a statement about a restricted subpopulation that shares *all* the individual's characteristics. Taken to extreme, such restrictive interpretation would insist on characterizing the plaintiff to minute detail, and would reduce the "but for"

probability to zero or one when all relevant details are accounted for. It is inconceivable that this interpretation underlies the intent of judicial standards. By using the wording "more probable than not," law makers have instructed us to ignore specific features which are either irrelevant or for which data are not likely to be available, and to base our determination on the most specific yet essential features for which data is expected to be available. In our example, two properties of Mr. A were noted: (1) that he died and (2) that he chose to use the drug; these are essential and were properly taken into account in bounding PN. In certain court cases, additional characteristics of Mr. A would be deemed essential. For example, it is quite reasonable that, in the case of Mr. A, the court may deem his medical record to be essential, in which case, the analysis should proceed by restricting the reference class to subjects with medical history similar to that of Mr. A. However, having satisfied such specific requirements, and knowing in advance that we will never be able to match *all* the idiosyncratic properties of Mr. A, the law makers' intent must be interpreted relative to the probability bounds provided by PN.

Conclusions

While reasoning from EoC to CoE involve the challenge of reasoning from group data to individual cases, the logical gulf between the two is no longer a hindrance to systematic analysis. It has been bridged by the structural semantics of counterfactuals (Balke and Pearl, 1994a,b) and now yields a coherent framework of fusing experimental and observational data to decide individual cases of all kinds, EoC included.

Glenn Shafer (2000) made an interesting observation in his essay on counterfactuals:

> "Even Laplace's vision of determinism, in which a superior but human-like intelligence can predict the future states of the world from knowledge of the present state and a small number of laws, demands only the possibility of prediction for states in which the world is actually found. If causal laws predict everything, they predict that the physician will undertake the operation. Thus the Laplacean vision does not require that the superior intelligence should be able to make a prediction about what would happen if the operation is not undertaken."

I believe Laplace would be surprised, and mighty gratified to know that his superior intelligence should be able to predict, not only what would happen if past actions were not undertaken, but doing so from statistical data on

past actions, without knowing the present state of the world nor the "small number of laws" that govern that world. This superior intelligence need only take seriously the Laplacean model and the counterfactual logic that it entails.

Acknowledgment

I am grateful to Nicholas Jewell and the editor of *Sociological Methods and Research* for calling my attention to the DFF's paper, and for helpful comments on the first version of the manuscript. Portions of this paper are based on Pearl (2000a, 2011, 2012). This research was supported in parts by grants from NSF #IIS1249822 and #IIS1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

Appendix A

This Appendix contrasts the results reported in this paper with the opinions expressed in Dawid et al. (2014a). For a more detailed comparison see (Pearl, 2014b).

DFF present the problem of CoE as a newly discovered challenge "for which the statistical literature is only of limited help." To remedy this neglect, they offer to "provide an alternative framing of the "CoE" that differ substantially from that found in the bulk of the scientific literature."

As part of this alternative framing, they propose $PC = P(R_0 = 0|R_1 = 1)$ as the parameter that need to be estimated for answering CoE questions. In their words:

> "To address the issue of whether taking the aspirin caused the observed recovery, we might ask: What is the probability that the (necessarily unobserved) potential response R_0, which would have been observed had I not taken aspirin ($A = 0$), would have been different ($R_0 = 0$) from that actually observed ($R_1 = 1$)."

The inappropriateness of PC as a measure of CoE was demonstrated in footnote 4. Here we simply note that the response "actually observed" is not $R_1 = 1$, but $R = 1$, and this flaw of formulation has had several repercussions on DFF conclusions.

In a discussion following DFF's paper, Nicholas Jewell alerted the authors to the "more relevant" interpretation of "but for" in terms of PN, to the extensive work done on CoE under this and other interpretations and, in particular, to the tight bounds derived by Tian and Pearl (2000) under a variety of assumptions, using both observational and experimental data (Jewell, 2014). In

their rejoinder (Dawid et al., 2014b), DFF explained their choice of the PC measure in these words:

> Jewell notes the close connection with earlier work of Robins and Greenland (1989; Greenland and Robins, 2000), and of Pearl and his collaborator Tian (Pearl, 2009; Tian and Pearl, 2000). We were aware of this work, having referenced it in earlier articles, and were remiss in not including discussion of it here. Robins and Greenland, using different notation and statistical formalisms, focus on what we and they call the PC although without the potential outcome labels, and they present the same lower bound, which come from the standard Fréchet bounds for 2×2 tables. They also address the assigned shares approach to interpreting the role of the relative risk used by the courts to address the CoE.
>
> Jewell suggests that we should have focused on $P(R_0 = 1|R_1 = 1$ and $A = 1)$ where A denotes the observed exposure condition– which is Pearl's Probability of Necessity (PN).This was in fact the way in which the CoE problem was initially formulated by Dawid (2011), the simplification to $Pr(R_0 = 1|R_1 = 1)$ being based on the "(questionable) assumption that the decision to take aspirin was unrelated to the (then hidden) values of the potential responses." Now this additional assumption is unreasonable unless the joint probability distribution being manipulated can be regarded as that fully specific to the given individual; and, to the extent that knowledge of this individual distribution is informed by EoC-type data, it will be essential that probabilities estimated from these data are computed relative to a suitably refined reference class. Without this requirement, focusing on bounds for $P(R_0 = 1|R_1 = 1$ and $A = 1)$ will not be the right thing to do.
>
> We also note that the difference in the condition for our PC and Pearl's is what led to the upper bound in Pearl's work with Tian, which is not necessarily 1 for PN. Moreover, the work of Pearl and others to sharpen these bounds and to identify PN rests on heroic assumptions that we deem inappropriate for the present discussion, especially when they ignore the distinctions between populations and samples, and observational and experimental data. Dawid et al. (2014b) do provide a more general treatment than the one we do in our article, which does allow for an upper bound that can differ from 1, but again it differs from that of Tian and Pearl for the reasons given previously.

In (Pearl, 2014b), I analyze the inconsistencies of these paragraphs, as well as the reasons why the "more general" analysis of Dawid et al. (2014b) has not produced the simple and informative bounds presented in Section 5 of this paper. Here we merely list the major opportunities missed in Dawid et al. (2014b).

1. Dawid et al. (2014b) failed to realize that fusing observational and experimental data can provide information that each study in isolation cannot, and that the fusion may produce a solution to the CoE problem.

2. Dawid et al. (2014b) failed to realize that confounding can be a blessing in that it may raise the lower bound (when $CF > 0$) and lower the upper bound (when $CF < 0$) (as shown in Figs. 1(a,b)).

3. Dawid et al. (2014b) treat the PN bounds as an evidence that CoE is a hard, if not metaphysical problem. They fail to appreciate the validity of claims based on these bounds. In particular, if the lower bound is above 50% then, the "more probable than not" criterion is met, and no further assumption (beyond the suitability of the reference class) is needed to substantiate this claim.

References

Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 46–54.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

Cai, Z. and Kuroki, M. (2006). Variance estimators for three 'probabilities of causation'. *Risk Analysis* **25** 1611–1620.

Cole, P. (1997). Causality in epidemiology, health policy, and law. *Journal of Marketing Research* **27** 10279–10285.

Dawid, A. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association* **95** 407–448.

Dawid, A. (2011). The role of scientific and statistical evidence in assessing causality. In *Perspectives on Causation* (R. Goldberg, ed.). Hart Publishing, Oxford, England, 133–147.

Dawid, A., Fienberg, S. and Faigman, D. (2014a). Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociological Methods and Research* **43** 359–390.

Dawid, A., Musio, M. and Fienberg, S. (2014b). From statistical evidence to evidence of causality. Tech. rep., Statistical Laboratory, University of Cambridge, UK. Submitted to *Bayesian Analysis*. ArXiv: 1311.7513.

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Freedman, D. A. and Stark, P. B. (1999). The swine flu vaccine and Guillain-Barré syndrome: A case study in relative risk and specific causation. *Evaluation Review* **23** 619–647.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

Greenland, S. (1999). Relation of probability of causation, relative risk, and doubling dose: A methodologic error that has become a social problem. *American Journal of Public Health* **89** 1166–1169.

Greenland, S. and Robins, J. (2000). Epidemiology, justice, and the probability of causation. *Jurimetrics* **40** 321–340.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Jewell, N. P. (2014). Assessing causes for individuals: Comments on Dawid, Faigman, and Fienberg. *Sociological Methods and Research* **54** 391–395.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

Pearl, J. (2000a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2000b). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association* **95** 428–431.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2011). The algorithmization of counterfactuals. *Annals for Mathematics and Artificial Ingelligence* **61** 29–39.

Pearl, J. (2012). The causal foundations of structural equation modeling. In *Handbook of Structural Equation Modeling* (R. Hoyle, ed.). Guilford Press, New York, 68–91.

Pearl, J. (2013). The curse of free-will and paradox of inevitable regret. *Journal of Causal Inference* **1** 255–257.

Pearl, J. (2014a). Causes of effects and effects of causes. Tech. Rep. R-431, <http://ftp.cs.ucla.edu/pub/stat_ser/r431.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Short version forthcoming, Journal of Sociological Methods and Research.

Pearl, J. (2014b). A note on causes of effects. Tech. Rep. R-439, <http://ftp.cs.ucla.edu/pub/stat_ser/r439.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Robertson, D. (1997). The common sense of cause in fact. *Texas Law Review* **75** 1765–1800.

Robins, J. and Greenland, S. (1989). The probability of causation under a stochastic model for individual risk. *Biometrics* **45** 1125–1138.

Rubin, D. (2014). Quoted in Li, F. and Mealli, F., "A conversation with Donald B. Rubin". Tech. rep., Department of Statistical Science, Duke University, Durham, NC. Http://arxiv.org/abs/1404.1789. Forthcoming, *Statistical Science*.

Shafer, G. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association* **95** 438–442.

Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (J. Bilmes and A. Ng, eds.). AUAI Press, Corvallis, OR, 514–521.

Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence* **28** 287–313.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.

Yamamoto, T. (2012). Understanding the past: Statistical analysis of causal attribution. *American Journal of Political Science* **32** 237–256, DOI: 10.1111/j.1540–5907.2011.00539.x.