ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Effects of Causes and Causes of Effects

A. Philip Dawid[1] and Monica Musio[2]

[1] Statistical Laboratory, Faculty of Mathematics, University of Cambridge, Cambridge CB3 0WB, United Kingdom; email: apd@statslab.cam.ac.uk

[2] Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, 09124 Cagliari, Italy; email: mmusio@unica.it

Annu. Rev. Stat. Appl. 2022. 9:261–87

First published as a Review in Advance on November 4, 2021

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-070121-061120

Keywords

causal Bayesian network, counterfactual, decision-theoretic causality, directed acyclic graph, instrumental variable, interval of ambiguity, potential outcome, probability of causation, statistical causality, stochastic causal model, structural causal model, structural equation model, twin network

Abstract

We describe and contrast two distinct problem areas for statistical causality: studying the likely effects of an intervention (effects of causes) and studying whether there is a causal link between the observed exposure and outcome in an individual case (causes of effects). For each of these, we introduce and compare various formal frameworks that have been proposed for that purpose, including the decision-theoretic approach, structural equations, structural and stochastic causal models, and potential outcomes. We argue that counterfactual concepts are unnecessary for studying effects of causes but are needed for analyzing causes of effects. They are, however, subject to a degree of arbitrariness, which can be reduced, though not in general eliminated, by taking account of additional structure in the problem.

1. INTRODUCTION

1.1. Overview

The enterprise of statistical causality has seen much activity in recent years, both in its foundational and theoretical aspects, and in applications. However, it remains rare to draw the distinction (recognized by Mill 1843) between two different problem areas within it: assessing (in individual cases, or in general) the likely effects of applied or considered interventions—the problem of effects of causes (EoC)—and assessing, in an individual case, whether or not an observed outcome was caused by an earlier intervention or exposure—the problem of causes of effects (CoE). Where this distinction is made, it is typically assumed that both problems can be represented and addressed using a common theoretical framework, such as the structural causal model of Pearl (2009).

Cuellar (2017) considered the general problem of assessing CoE and identified four stages, which apply equally to assessing EoC:

1. How should we define the probability of interest?
2. How should we identify this probability (and on the basis of what assumptions)?
3. How should we estimate this probability?
4. How should we use the results of the analysis?

The present article largely concentrates on stages 1 and 2. Our purpose is to emphasize the important logical and technical differences between EoC and CoE problems and to explore and compare the various ways in which problems of each kind can be and have been formulated. In particular, we argue that different tools are appropriate for the two different purposes.

In Section 1.2, we introduce the variety of concerns to be addressed, in the context of a specific lawsuit; we expand on these in Section 1.3. In Section 2, we introduce and compare a variety of formalisms that have been proposed to address EoC. Then, we briefly summarize some philosophical and implementational issues. Section 2.4 introduces, with examples, the problem of inference in the presence of an instrumental variable, which is then used throughout Section 2 as a hook on which to hang the general discussion. Section 2.5 describes purely probabilistic aspects. Then Section 2.6 introduces the decision-theoretic (DT) approach to EoC, Section 2.7 an approach based on linear models, Section 2.8 a nonparametric generalization of that, and Section 2.9 the approach based on potential outcomes.

We turn to address CoE in Section 3 for problems similar to those of Section 1.2. Section 3.1 points to the need for counterfactual inference, which cannot, however, totally resolve the ambiguities inherent in such problems. Two ways of conducting counterfactual modeling, based on structural and stochastic causal models, are described in Section 3.2, and we show how they can both be subsumed in the potential outcome approach of Section 3.3. In Section 3.4, we consider how empirical data can be used to inform CoE analysis but cannot totally resolve the inherent ambiguities. In Section 3.5, we address the legal CoE issues of Section 1.2, showing how the basic ambiguity, expressed by interval bounds on the probability of causation (PC), can be refined when we can observe other variables in the problem. Section 3.6 indicates just how limited our CoE analyses have been and what difficulties might attend further extension.

To conclude, Section 4.1 summarizes some of the lessons to be learned from our review of the different approaches to EoC and CoE, while Section 4.2 revisits the lawsuit of Section 1.2, comparing our analysis with legal opinions.

1.2. A Lawsuit

In 2014 a class action (multidistrict litigation) was brought in the United States by more than 3,000 women who sued the pharmaceutical company Pfizer, claiming that they developed (type 2) diabetes as a result of taking its drug Lipitor (atorvastatin calcium) (*Hempstead v. Pfizer, Inc.* 2015). The plaintiffs identified two bellwether cases of women making such a claim for closer attention.

In order to succeed in such a suit, the plaintiffs would have to demonstrate, in succession, to the legal system's satisfaction, two points (Dawid et al. 2014, Sanders et al. 2021):

- General causation: Can Lipitor cause diabetes?
- Specific causation: In the individual cases, did Lipitor cause their diabetes?

The eventual judgment was in favor of Pfizer. It was judged that general causation had not been established for doses of 10, 20, and 40 mg of the drug but could be considered for the 80-mg dose. And with regard to the bellwether cases, it was judged that specific causation could not be established.

The distinction the court made between the two varieties of causal question, general and specific, is fundamental and occurs in many contexts. It has various descriptions. Philosophers talk about type and token causation. Legal scholars talk about group and individual causation and have coined the expression G2i (group to individual) for the task of arguing from one to the other (Faigman et al. 2014). In statistical contexts we may talk about inference about effects of causes, and causes of effects, which are the designations we mostly use here.

1.3. Statistical and Causal Questions

Questions about individual cases can usefully be organized in a fourfold classification.[1] We exemplify these for the bellwether case of Juanita, who is 55 years old and has a total cholesterol of 250 mg/dL, low-density lipoprotein of 175 mg/dL, high-density lipoprotein of 46 mg/dL, triglyceride level of 142 mg/dL, weight of 176 lbs, and body mass index of 26.37.

- Forecasting: Juanita has started taking a 80 mg dose of Lipitor daily. Is she likely to develop diabetes?
- Backcasting: Juanita has developed diabetes. Did she take Lipitor, and if so, in what dose and for how long?
- Decision: Juanita is considering whether to take Lipitor but is worried about developing diabetes. What should she do?
- Attribution: Juanita took Lipitor 80 mg daily for 3 years and developed diabetes. Was that because she took Lipitor?

While forecasting and backcasting are fundamentally purely statistical exercises, decision and attribution can be classified as causal questions—the former addressing EoC and the latter addressing CoE.

1.3.1. Forecasting.

Forecasting is an apparently straightforward statistical task, at least conceptually: We gather high-quality data on individuals sufficiently like Juanita, taking the same treatment, and observe the proportion who go on to develop diabetes. In practice this simple recipe will be complicated by nonrandom sampling of cases, differences in background characteristics, difficulties associated with long-term follow-up, censoring by death, and so on. Handling such complications has been a prime focus of statistical research over many decades, and though the issues raised are very far from trivial, they raise no new issues of principle. But we would also need to argue that the proportion, estimated from the data, of individuals developing diabetes can

[1] The distinction between forecasting, backcasting, and attribution was made by Honoré (2010) in a legal context and by Dawid (2013) in a statistical context.

be identified with Juanita's individual risk. While this does raise some subtle philosophical issues (Dawid 2017), they can largely be ignored for practical purposes.

1.3.2. Backcasting.

Backcasting refers to the task of predicting uncertain past events on the basis of later observations. In a statistical context, this is most typically performed by application of Bayes' theorem. Suppose we do not know whether or not Juanita took the Lipitor but, as above, have estimated the two forward forecast probabilities under each scenario. We would also need to assign a prior probability to the event that she did, in fact, take the drug. Bayes' theorem supplies the machinery for combining these ingredients to produce the required backward probability that she indeed took Lipitor, on the basis of her having developed diabetes. Although such Bayesian inferences have, from the very beginning, often been described as estimating the probabilities of causes, use of the term cause here is not really appropriate, since even if we can conclude that Juanita had taken Lipitor, that might not have been the cause of her diabetes.

Applying Bayes' theorem is not the only way to conduct backcasting. More straightforwardly, we could simply collect a sample of individuals sufficiently like Juanita, confine attention to those who develop diabetes, and use the proportion of these who had taken Lipitor to estimate the desired probability for Juanita. Indeed, there are circumstances where this simple approach may be preferable to the Bayesian route (Dawid 1976).

1.3.3. Decision.

Forecasting is of fundamental importance in decision analysis. Suppose Juanita has not yet started taking Lipitor, and is considering whether or not to do so. One of her concerns is whether she will develop diabetes. She should thus consider and compare how probable this event is under two possible scenarios: that she does, or that she does not, take the drug. This would require two separate forecasting exercises and, correspondingly, data from two different sets of individuals who do or do not take Lipitor.

But new difficulties now arise in gathering and using such data. In particular, the very treatment desired by such an individual might be related to her overall health status and thus affect her risk of developing diabetes—even were she not to receive that desired treatment. In such a case it becomes problematic to disentangle the effects of desire for treatment and of application of treatment. This is an example of the problem of confounding, which requires careful attention in such cases.

1.3.4. Attribution.

Questions of forecasting, backcasting, and decision, although beset with many practical difficulties, can all, in principle at least, be answered directly by means of probabilities attached to unknown events of interest—probabilities that can be estimated given suitable data. However, a question of attribution, such as "Did taking Lipitor cause Juanita's diabetes?" is not so readily resolved, for what is it now that is unknown? We know that Juanita took Lipitor, and we know that she developed diabetes. There is no unknown event about which we require inference. Rather, it the relationship between these events that is uncertain—was it causal or not? Even to understand what we might mean by such a question is problematic.

We consider how to formalize such questions and explore just what can be concluded from data about them in Section 3.

2. EFFECTS OF CAUSES

2.1. Causality and Agency

Philosophers, on the one hand, have debated causality for millennia and have propounded a large variety of conceptions and approaches. Statisticians, on the other hand, had traditionally been reluctant to imbue their inferences with causal meaning. But in recent years much more attention has been given to what we can now term statistical causality. Particularly influential have been the contributions of Rubin (1974), who promoted a formulation based on potential outcomes, and of Pearl (2009), based on graphical representations.

Implicit in both these approaches is the idea of a cause as an intervention applied to a system, in line with the agency interpretation of causality (Reichenbach 1956; Price 1991; Hausman 1998; Woodward 2003, 2016). A main task for statistical causality is to make inference about the effects of such interventions—that is, understanding the EoC—on the basis of data. When making use of data, it is important to distinguish between data generated through experimentation and purely observational data.

2.2. Experiment

In an experiment, interventions are made on experimental units according to some known protocol, often involving randomization, and their responses measured. To the extent that the experimental units and interventions can be regarded as representative of future interventions on new units, it is in principle straightforward to infer what effects those interventions will have in future. Design and analysis of experiments is a major enterprise within modern statistics, involving many subtle and technical considerations, but no special issues of principle arise.

2.3. Observation

Things are not so straightforward when the data available are purely observational and the process whereby treatment interventions were applied to units is not known. For example, when choosing between two treatments, a doctor may have given one preferentially to those patients he considers sicker. Then a simple comparison of the outcomes in the two treatment groups will be misleading, since even if there is no difference between the treatments, a difference in outcomes may be seen because of the difference in general health of the two treatment groups. This is the problem of confounding, which prevents us from taking the observational data at face value. In such a case, it may or may not be possible to assess, by more sophisticated means, genuine causal effects, depending on what is observed and what assumptions can reasonably be made. If we know or can reasonably assume how the doctor behaved, and we have data on the patient characteristics that the doctor used, then we can make meaningful comparisons and extract causal conclusions, but—in the absence of further structure or assumptions—this will not be the case if either of the conditions fails.

Much of the modern enterprise of statistical causality is focused on addressing this issue of extracting causal conclusions from observational data. In order to do so, it will invariably be necessary to make assumptions, which are generally untestable in practice, about the relationship between the behaviors of the idle observational system, which generates the observed data, and the same system under a specified intervention—which is what is wanted, but is not directly observed. Such assumptions are sometimes made explicit, and so open to reasoned scrutiny and debate, but sometimes they remain implicit and hidden, taken for granted without critical examination. The kind of relationships required can typically be expressed, explicitly or implicitly, as asserting the equality of certain ingredients in both idle and interventional circumstances. While such invariance properties have sometimes been taken as the very definition of causality (Bühlmann 2020), they can be applied without any such philosophical commitment. (Our own philosophical standpoint remains that based on agency.)

The do-calculus (Pearl 2009, section 3.4; see also Dawid 2015, section 9.7) applies to problems that can be modeled by means of a directed acyclic graph (DAG) representing both assumed conditional independence properties of the observational regime and assumed relationships between the observational and interventional regimes. For such a case, do-calculus supplies a complete method for determining whether a causal estimand of interest can be identified from observational data and, if so, how.

2.4. Instrumental Variable

Below, we introduce, compare, and contrast some of the different statistical formalisms that have been used to model EoC. To be concrete, we consider how each formalism might model an instrumental variable problem (Bowden & Turkington 1984). This involves, in addition to the treatment variable X and response variable Y, a further observed variable Z (the instrument) and an unobserved variable U—all defined for individuals in a study or larger population. Typically Z is binary, X and Y are binary or continuous, and U is multivariate. Note that in this problem it is not possible, without imposing still further structure, to identify the causal effect of X on Y from observational data.

We suppose the following: (*a*) U is a set of preexisting characteristics of the individual. (*b*) Z is associated with X but not with U. (*c*) While X could in principle be assigned externally, in the study it was not. (*d*) Given X, and the individual characteristics U, the response Y is unaffected by Z. (This vague requirement, also called the exclusion restriction, is clarified further below.)

Example 1 (Encouragement trial). In an encouragement trial (Holland 1988), students are randomly assigned to receive encouragement to study. However, a student may or may not respond to the encouragement. Here Z is a binary assignment indicator, taking value 1 for encouragement, 0 for no encouragement; X is the number of hours the student actually studies; Y is the student's score in the final test; and U comprises individual characteristics of the student that may affect both X and Y. Because of randomization, Z is independent of U. We are interested in how a student's choice of study hours affects their test score.

Example 2 (Incomplete compliance). In a medical trial, each patient is randomly assigned to take either active treatment ($Z = 1$) or a placebo ($Z = 0$). However, the patient may not comply with the assignment, so that the treatment actually taken, $X = 1$ or 0, may differ from Z. Finally, we observe whether the patient recovers ($Y = 1$) or not. We allow for possible dependence of both X and Y on further unobserved patient characteristics, U. Again, randomization ensures that Z is independent of U. We are principally interested in the effect of taking the treatment on recovery.

Example 3 (Availability trial). In a variation of Example 2, $Z = 1$ means the active treatment is made available to the patient, and $Z = 0$ means that it is not; $X = 1$ if the treatment is taken, and $X = 0$ if it is not. It is supposed that if the treatment is unavailable ($Z = 0$) it cannot be taken ($X = 0$) (though it need not be taken when it is available).

Example 4 (Mendelian randomization; Katan 1986). Low serum cholesterol level ($X = 1$) is thought to be a risk factor for cancer ($Y = 1$). Both serum cholesterol and cancer may be affected by indicators of lifestyle (U). Possession of the E2 allele ($Z = 1$) of the apolipoprotein E (APOE) gene is known to be associated with low serum cholesterol level; this relationship need not be causal, but may arise because APOE is in linkage disequilibrium with the actual causative gene. Since nature randomizes the APOE allele at birth, and its level is not thought to affect lifestyle, U should not be associated with Z. We are interested in whether intervening to raise serum cholesterol could lower the risk of cancer.

There are a number of questions we could ask (but would not necessarily be able to answer) in such examples. In Example 2, these might include the following:

1. What is the probability of recovery for a patient who is assigned to active treatment (irrespective of the treatment actually taken)?
2. What is the probability of recovery for a patient who (irrespective of assigned treatment) in fact took active treatment?
3. What is the probability that a patient who recovered complied with the assignment?
4. What is the effect on recovery of assignment to treatment?
5. What is the causal effect of taking the treatment on recovery?

Questions 1, 2 and 3 inhabit the lowest rung, seeing, of Pearl's ladder of causation (Pearl & Mackenzie 2018), the first two being instances of forecasting, and the last of backcasting. Questions 4 and 5 are on the second rung of the ladder, doing, being instances of decision. In the following, we are mainly interested in 5. Since Z has been randomized, we could argue that the intention-to-treat question 4 is essentially the same as question 1, which can be straightforwardly addressed from the study data on (Z, Y). However, question 5 cannot readily be answered in the same way as question 2 (an as-treated analysis) since X has not been randomized, and any observed association between X and Y might be due to their common dependence on U.

2.5. Seeing: Conditional Independence

We first consider how to express, formally, purely probabilistic properties of the observational joint distribution of (X, Y, Z, U). This is all that is required to address forecasting and backcasting questions such as questions 1–3. However, it is not possible to formulate, let alone solve, causal queries such as question 5 in this setting: These live on the second rung, doing.

Specifically, assumption *b* implies that Z is independent of U: Using the notation of Dawid (1979), we write this as

$$Z \perp\!\!\!\perp U. \qquad 1.$$

Here, assumption *d* is interpreted as asserting the probabilistic independence of Y from Z, conditional on X and U:

$$Y \perp\!\!\!\perp Z \mid (X, U). \qquad 2.$$

2.5.1. Graphical representation.

It is often convenient to display such conditional independence properties by means of a DAG. Each node in the DAG represents a variable in the problem, and missing arrows represent assumed properties of conditional independence in their joint distribution—Dawid (2015, section 6) provides full details. The DAG is a partial description, displaying only qualitative aspects, of the joint distribution.

The DAG representing Property 1 and Property 2 looks like **Figure 1** (the dotted outline of U is nonessential, merely a reminder that U is unobserved). The absence of an arrow between Z and U represents their independence (Property 1), while the missing arrow from Z to Y represents their conditional independence given the parents of Y, namely X and U (Property 2). In general, in a DAG representation, any variable is conditionally independent of its nondescendants, given its parents. Further conditional independence properties implied by these can be read off the DAG, using the *d*-separation (Verma & Pearl 1990) or equivalent moralization (Lauritzen et al. 1990) criteria, as described by Dawid (2015).

The qualitative DAG representation of a joint distribution can be expanded to a full quantitative description by specifying, for each variable, its conditional distribution given its parents in the DAG. (This would be required to encode the condition in assumption *b* that X is not independent

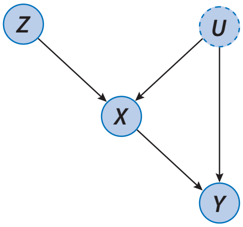

Figure 1

Instrumental variable: seeing. The graph represents purely probabilistic properties, specifically that Z and U are independent, and that Y is independent of Z, given X and U.

of Z.) Elegant algorithms exist, taking advantage of the DAG structure, for streamlining quantitative computation of joint and conditional probabilities (Cowell et al. 1999). Such probabilities are what is needed to address questions of forecasting and backcasting.

2.6. Doing: Decision-Theoretic Causality

The DT approach to causality has been described in this journal (Dawid 2015); its foundational underpinnings are examined by Dawid (2021).

We have several regimes of interest. For each possible value x of X, we have an interventional regime, where treatment value x is forced on an individual (that is, X is set to x, which we notate as $X \leftarrow x$). We also have an idle regime, in which the treatment X is merely observed, and any value may occur. It is helpful to introduce a nonstochastic regime indicator variable F_X, where $F_X = x$ labels the interventional regime with $X \leftarrow x$, and $F_X = \emptyset$ labels the idle regime. The response variable Y may have different distributions in the different regimes. The object of causal inference will usually be some contrast between the response distributions in the various interventional regimes—this is what is required to address the decision problem of choosing which value to set X to. For example, when X is binary, interest typically centers on the difference in the expected response between the two interventional regimes, $\mathrm{E}(Y \mid F_X = 1) - \mathrm{E}(Y \mid F_X = 0)$, which is termed the average causal effect (ACE).

But in the cases to be considered, there are no data available that are directly relevant to the interventional settings of interest, and we want to make use of observational data collected under the idle regime, $F_X = \emptyset$, to make inferences about what would happen in interventional settings. This may or may not be possible. At the least, it is necessary to make, and justify, relationships between the idle and the interventional settings. DT studies when and how such relationships can be used to support causal inference from observational data.

For example, we might be willing to assume, in addition to Assumptions 1 and 2, that no matter whether X is merely observed ($F_X = \emptyset$) or is set by external intervention ($F_X = 0$ or 1), the following ingredients will be the same: (*e*) the distribution of Z; (*f*) the distribution of U, with U independent of Z; and (*g*) the conditional distribution of Y given (Z, X, U) [which would then in all cases depend only on (X, U), since this is so under regime $F_X = \emptyset$, by assumption *d*].

These properties do not follow logically from Properties 1 and 2 (Dawid 2022), and if they are to be applied they need additional arguments, as described by Dawid (2021).

We can interpret assumptions *e*–*g* as conditional independence properties:

$$Z \perp\!\!\!\perp F_X, \qquad 3.$$

$$U \perp\!\!\!\perp (F_X, Z), \text{ and} \qquad 4.$$

$$Y \perp\!\!\!\perp (F_X, Z) \mid (X, U). \qquad 5.$$

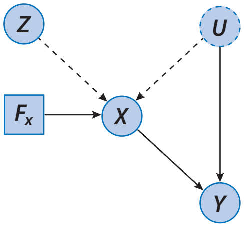

Figure 2

Instrumental variable: doing. The graph of **Figure 1** has been augmented with the regime indicator F_X. It represents additional causal properties: that U and Z have the same joint distribution, both in interventional and in observational regimes, and that the distribution of Y given Z, X, and U is the same in all regimes.

Even though F_X is a nonstochastic indicator of regime (observational/interventional), these intuitively meaningful extended conditional independence expressions can be manipulated essentially just as if F_X were a random variable (Constantinou & Dawid 2017).

In this approach, the conditions *e–g*, or equivalently Properties 3–5, which relate to behavior under possible intervention at X, are the full causal ingredients of our model.

2.6.1. Graphical representation. We can augment **Figure 1** with an additional node for F_X (square to indicate it is nonstochastic), and we then obtain **Figure 2**. This DAG represents (in exactly the same way as before) the assumed conditional independencies of Properties 1–5, which fully embody our causal assumptions. [The dashed arrows from Z and U to X are there to indicate that they are absent under an interventional regime $F_X = x$, since then we have $X = x$, trivially independent of (Z, U).] Note in particular that the arrow from Z to X in **Figure 2** does not encode a causal effect of Z on X, since assumptions *e–g* are fully consistent with cases, such as Example 4, where Z and X are merely associated (Dawid 2010, section 10).

2.6.2. Causal Bayesian network. Pearl (2009) uses the same causal semantics as described above to construct what he terms a causal Bayesian network (CBN). The difference is that he would normally consider the possibility of intervention on every observable variable, which, in our case, would mean adding further intervention indicator nodes F_Z and F_Y to **Figure 2**—parents, respectively, of Z and Y. In such a case, the presence of all the intervention nodes is usually taken for granted and omitted from the augmented DAG, rendering it visually indistinguishable from an unaugmented DAG (**Figure 1**). However, there are clear advantages to retaining explicit intervention nodes in the figure:

1. This eliminates the possibility of confusion between rung 1 (seeing) and rung 2 (doing) interpretations of apparently identical DAGs.
2. The causal links assumed between regimes are fully represented by *d*-separation properties of the augmented DAG.
3. It will (as above) often be appropriate to consider interventions on only some of the variables. In particular, there will then be no need to impose the additional cross-regime causal constraints associated with further, inessential intervention indicators.

2.6.3. Estimation. Even after assuming links, as above, between the observational and interventional regimes, it does not follow that we have enough structure to enable us to use the observational data to estimate, say, the causal effect, ACE, of X on Y. And indeed, in this example, further structure must be imposed to support such causal inference. For instance, in Example 1 we might require that Y has a linear regression on (X, U) (this being the same in all regimes, by Property 5):

$$\mathrm{E}(Y \mid X, U, F_X) = W + \beta X, \tag{6.}$$

where W is a function of U. Since then $\mathrm{E}(Y \mid F_X = x) = w_0 + \beta x$, where $w_0 = \mathrm{E}(W)$, β has a clear causal interpretation. Also, restricting attention to the observational regime $F_X = \emptyset$, we obtain $\mathrm{E}(Y \mid Z) = \mathrm{E}\{\mathrm{E}(Y \mid X, U, Z) \mid Z\} = \mathrm{E}\{\mathrm{E}(Y \mid X, U) \mid Z\}$, by Property 5, $= w_0 + \beta \mathrm{E}(X \mid Z)$, by Property 4. This implies that we can estimate β, from the observational data, as the ratio of the coefficients of Z in the sample linear regressions of Y on Z and of X on Z.

In cases such as Example 2 with binary X, Equation 6 is equivalent to

$$\mathrm{SCE}(u) = \beta, \tag{7.}$$

a constant for all u, where $\mathrm{SCE}(u) = \mathrm{E}(Y \mid U = u, F_X = 1) - \mathrm{E}(Y \mid U = u, F_X = 0)$ is the specific causal effect of X on Y, relevant to the subpopulation having $U = u$. That is to say, the specific causal effect is required to be nonrandom, the same in all subpopulations. Then ACE $= \mathrm{E}\{\mathrm{SCE}(U)\} = \beta$ also, and so is estimable as above. Alternatively, when all variables are binary, without making any modeling assumptions we can determine bounds on ACE from the data (Balke & Pearl 1997, Dawid 2003).

2.7. Linear Structural Equation Model

Linear structural equation modeling (SEM), closely related to path analysis (Wright 1921), is perhaps the earliest approach to instrumental variable problems—and much else besides. It can be considered as an extension of linear regression modeling. In the context of the encouragement trial of Example 1, we might express the relationship between Z, X, and Y by the pair of regression-like equations:

$$X = \alpha_0 + \alpha_1 Z + U_X, \tag{8.}$$

$$Y = \beta_0 + \beta_1 X + U_Y. \tag{9.}$$

(Such a system would often be completed with a further equation for Z, which here would simply be $Z = U_Z$. However, we omit this on account of its triviality.) Here, U_X and U_Y are zero-mean residual error terms. In this problem it would be assumed that U_X and U_Y are uncorrelated with Z but not necessarily with each other. The absence of Z in Equation 9 embodies the exclusion restriction.

This model can be rendered graphically as in **Figure 3**, which may be compared with **Figure 1**, identifying $U = (U_X, U_Y)$.

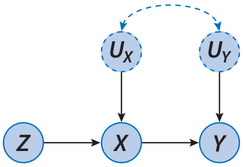

Figure 3

Structural equation graph. U_X and U_Y are correlated random disturbances. X is a linear function of random Z and U_X. Y is a linear function of X and U_Y. This graph can be given a causal, as well as a purely probabilistic, interpretation.

As discussed by Pearl (2009, section 5.1.2), the intended interpretation—in particular, the causal interpretation—of SEM has often been unclear. Pearl's suggestion is as follows:

1. In the system of Equations 8–9, X is functionally determined by (Z, U_X), and Y is functionally determined by (X, U_Y). So we can solve for (X, Y) in terms of (Z, U_X, U_Y). If we have a joint distribution for (Z, U_X, U_Y), this determines a joint distribution for (Z, X, Y), which can be regarded as representing the undisturbed system.
2. If, alternatively, we intervene to set X to x, it is assumed that we can replace Equation 8 by $X = x$ but retain Equation 9 essentially as is, so that $Y = \beta_0 + \beta_1 x + U_Y$, with the distribution of U_Y unchanged.

This approach gives a causal semantics to an SEM, relating the observational regime with possible interventional regimes. As with any such assumed relationship, it is not to be taken for granted, but argued for in the context of each particular problem. When we can assume condition 2, β_1 has a clear causal interpretation, being the rate of change of $\mathrm{E}(Y \mid X \leftarrow x)$ with respect to the value set, x. However, since U_Y is correlated with U_X, and hence in general with X, β will not be the coefficient of X in the observational regression of Y on X and so cannot be identified from that. Instead we can argue that, since U_X is uncorrelated with Z, $\mathrm{E}(X \mid Z) = \alpha_0 + \alpha_1 Z$; and, since U_Y is uncorrelated with Z, $\mathrm{E}(Y \mid Z) = \beta_0 + \beta_1 \mathrm{E}(X \mid Z) = \beta_0 + \alpha_0\beta_1 + \alpha_1\beta_1 Z$. It again follows, as in Section 2.6.3, that β_1 can be identified as the ratio of the coefficients of Z in the observational regressions of Y on Z and X on Z.

2.8. Structural Causal Model

We can generalize the system of Equations 8–9 by dropping the linearity requirement, yielding

$$X = f_X(Z, U_X), \qquad 10.$$

$$Y = f_Y(X, U_Y), \qquad 11.$$

where f_X and f_Y are specified general functions of their arguments, and (U_X, U_Y) have a specified joint distribution, typically not independent of each other but jointly independent of Z. The absence of Z as an argument of f_Y embodies the exclusion restriction. Pearl & Mackenzie (2018) refer to such a nonparametric structural equation model as a structural causal model (SCM), and we use this designation in the following.

Again, this system can be represented graphically by **Figure 3**. But, with no loss of generality, we can use U instead of the pair (U_x, U_Y) and write the system as

$$X = f_X(Z, U), \qquad 12.$$

$$Y = f_Y(X, U), \qquad 13.$$

with U independent of Z. This system again determines a joint observational distribution for (X, Y, Z), which is represented by **Figure 1**. We might again imbue this structural equation system with causal semantics (whose relevance in a real-life context will need justification): Assume that, under an intervention $X \leftarrow x$, we can replace Equation 12 by $X = x$ and Equation 13 by $Y = f_Y(x, U)$, where U is supposed to retain its original distribution. The extended structure is then, again, represented by **Figure 2**. In particular, Y will be independent of Z in an interventional regime, where the dotted arrows in **Figure 2** are absent. In contrast, in the observational regime, the distribution of Y, given $X = x$ and $Z = z$, is that of $Y = f_Y(x, U)$ given $f_X(z, U) = x$; because of this conditioning, the value of Z will typically make a difference to the distribution of Y.

In using **Figure 1** (and, implicitly or explicitly, **Figure 2**) as representations of the SCM system of Equations 12–13, we are supplying these figures with yet another semantic interpretation, where the dependence of X and Y on their parents is taken as deterministic, not stochastic. This is to be contrasted with the CBN interpretation of Section 2.6.2, in which all relationships are allowed to be stochastic.

It can be shown that, by suitable choice for the distribution of its U (which distribution is, however, not uniquely determined), the SCM can fully reproduce the joint distribution of (X, Y, Z), in all regimes, implied by a given fully stochastic DT model. This property holds in general for any problem represented by a DAG. For identifying EoC, we gain nothing by replacing a stochastic DT model with a deterministic SCM.[2] In particular, we again cannot identify the causal effect of X on Y without further assumptions, such as linearity in Equation 13, as for Equation 9 (with U_Y some function of U).

2.8.1. A comment.

If, in an SEM or SCM, U is regarded as a persistent attribute of an individual, the assumed determinism would mean that we would get the same output each time we applied the same intervention to that individual. That would be an unreasonable assumption in most contexts. Consequently, we should normally consider U as also incorporating information specific to the occasion of application (including, perhaps, random error), varying from occasion to occasion. Nevertheless, assuming the distribution of U does not change, ACE will still be constant, and so be meaningful, across occasions.

2.9. Potential Outcomes

In the potential outcomes (PO) approach to statistical causality (Rubin 1974), for each possible value x of the treatment X, we conceive of a version $Y(x)$ of the outcome variable Y, with all of these versions coexisting, even before application of treatment. It is supposed that $Y(x)$ [or, to be more explicit, $Y(X = x)$] is the outcome that would be observed in the interventional regime $F_X = x$. Typically it is further assumed (and termed consistency) that in the idle regime $F_X = \emptyset$ also, whenever $X = x$, the outcome will be $Y = Y(x)$ [which is why we do not distinguish between $Y(X = x)$ and $Y(X \leftarrow x)$]. Consistency is required to relate the observational and interventional regimes.

In the special but common and important case of binary X,[3] intervening to set X to 1 would reveal the value of $Y(1)$, while $Y(0)$ would remain unobserved, with a similar result when interchanging 1 and 0. In this approach, the single response Y is replaced by a bivariate quantity $\mathbf{Y} = (Y(0), Y(1))$, which must thus be endowed with a bivariate distribution. The fundamental causal contrast, comparing the effects of the two interventions, is considered to be the individual causal effect (ICE), ICE $= Y(1) - Y(0)$. However, direct inference about ICE is complicated by the fact [termed the fundamental problem of causal inference by Holland (1986)] that, because it is logically impossible to intervene on the same individual in two mutually exclusive ways simultaneously, we can never observe ICE or estimate its distribution. For this reason, it is customary to

[2] This point is further discussed in Section 2.9.1 and, especially, Section 2.9.2. Also, while Balke & Pearl (1997) make essential use of the deterministic functional relationships in the SCM to derive estimable bounds on ACE in the case of binary variables, Dawid (2003) shows how the same bounds can be obtained from the purely stochastic DT model.

[3] Of course, similar considerations apply more generally.

divert attention to the expected individual causal effect, E(ICE). By linearity of expectation, this is $\mathrm{E}\{Y(1)\} - \mathrm{E}\{Y(0)\}$, each term of which involves only one intervention. This then is the PO version of ACE, as introduced in Section 2.6, with essentially the same interpretation. Note, however, that there is no analogue of ICE in the DT approach; neither is there any DT analogue of, say, var(ICE), which involves the correlation between $Y(0)$ and $Y(1)$—a correlation that can never be estimated, on account of the fundamental problem of causal inference.

If we start with an SCM representation of a system, we can use it to construct associated potential outcomes. For example, starting from Equation 12 and Equation 13, just define, for each z, $X(Z = z) = f_X(z, U)$ and, for each x, $Y(X = x) = f_Y(x, U)$. Under the SCM causal semantics, $X(Z = z) = f_X(z, U)$ is assumed to supply the value of Y when $X = x$, whether or not there are interventions at X or anywhere else in the system (except at Y itself); this corresponds to the PO consistency property. This construction of POs makes them all functions of U, whose distribution thus generates a joint distribution for all POs.

Typically, however, a PO analysis would not make explicit use of an exogenous variable such as U and might not want to require that there exist any real-world variable or set of variables U with the properties assumed in Section 2.8 (for example, that Y is fully determined by X and U). Instead, one starts by introducing, as primitives, jointly distributed stochastic potential outcomes, $X(Z = z)$, $Y(X = x)$, and $Y(Z = z)$, for all possible values of x and z, and working directly with them. The exclusion restriction now becomes $Y(Z = z) = Y\{X = X(Z = z)\}$.

Now, introduce a new variable V, which is simply the collection of all $X(Z = z)$s and $Y(X = x)$s, as z and x vary. Then X is fully determined by (Z, V): When $Z = z$, we simply select the relevant element $X(Z = z)$ of V (this being valid in all regimes, by consistency); similarly, Y is determined by (X, V). We thus recover a formal[4] identity with the SCM of Equations 12 and 13, with V substituting for U—so long as we have V independent of Z. If we were starting from an SCM, as above, the independence of U and Z, and thus of V and Z, would be easy to justify since U represents preexisting characteristics of the individual and Z is randomized. Making a similar argument for V when taking POs as primitive is more problematic, since V does not correspond to any real-world quantity [in particular, on account of the fundamental problem of causal inference, certain elements of V, e.g., $X(Z = 1)$ and $X(Z = 2)$, are not simultaneously observable]. Nevertheless, the typical assumption is that it is indeed meaningful to consider the collection V of all possible potential responses as a preexisting (albeit unobservable) characteristic of the individual and thus argue that V is independent of the randomized variable Z. In this case, we recover a purely formal identity with an SCM. Now the linearity condition Equation 6 is equivalent to $Y(X = x) - Y(X = x') = \beta(x - x')$, which is thus required to be nonrandom.

2.9.1. A variation.

The above specifications can be considered simply as more detailed ways of realizing the CBN structure of Section 2.6, which is more general since in a CBN we need not assume the existence of potential outcomes, which cannot be derived from its stochastic form. And indeed, for estimating the ACE, the extra structure imposed beyond that of a CBN does not offer any improvement. But an SCM or PO approach allows us to formulate, and purports to solve, other causal questions. We consider one such in the context of Example 2, where Z, X, and Y are all binary (Imbens & Angrist 1994, Angrist et al. 1996).

[4] In a genuine SCM, U is regarded as a set of unobserved real-world background variables, with an appropriate, in principle knowable, distribution that, together with X, would determine Y. But it is hard to conceive of such a real-world interpretation of V.

Let $\mathbf{X}$ denote the pair $[X(Z=0), X(Z=1)]$ and $\mathbf{Y}$ the pair $[Y(X=0), (Y(X=1)]$. We assume consistency, the exclusion restriction $Y(Z=z) = Y\{X = X(Z=z)\}$, and that

$$Z \perp\!\!\!\perp (\mathbf{X}, \mathbf{Y}). \qquad 14.$$

It is easy to see that

$$Y(Z=z) = Y(X=1)X(Z=z) + Y(X=0)\{1 - X(Z=z)\}. \qquad 15.$$

We can define the following ICEs:

$$\mathrm{ICE}_{Z\to X} = X(Z=1) - X(Z=0), \qquad 16.$$

$$\mathrm{ICE}_{Z\to Y} = Y(Z=1) - Y(Z=0), \text{ and} \qquad 17.$$

$$\mathrm{ICE}_{X\to Y} = Y(X=1) - Y(X=0), \qquad 18.$$

and we can deduce from Equation 15 that

$$\mathrm{ICE}_{Z\to Y} = \mathrm{ICE}_{Z\to X} \times \mathrm{ICE}_{X\to Y}. \qquad 19.$$

We note that, since Z is randomized, $\mathrm{ACE}_{Z\to X} = \mathrm{E}(\mathrm{ICE}_{Z\to X}) = \mathrm{E}(X \mid Z=1) - \mathrm{E}(X \mid Z=0)$ is readily estimable from the observational data, and so, likewise, is $\mathrm{ACE}_{Z\to Y} = \mathrm{E}(Y \mid Z=1) - \mathrm{E}(Y \mid Z=0)$. However, there is no immediate parallel for $\mathrm{ACE}_{X\to Y}$, since X has not been randomized.

If we could replace each ICE term in Equation 19 by its expectation ACE, we would have

$$\mathrm{ACE}_{X\to Y} = \mathrm{ACE}_{Z\to Y}/\mathrm{ACE}_{Z\to X}, \qquad 20.$$

where the right-hand side of Equation 20 is estimable from the observational data (it is assumed that Z has a causal effect on X, so that $\mathrm{ACE}_{Z\to X} \neq 0$).

When we can assume, as for Equation 7, that $\mathrm{ICE}_{X\to Y} = \beta$ is nonrandom, we can take expectations in Equation 19, and Equation 20 does indeed hold, allowing estimation of β. But more generally, $\mathbf{X}$ and $\mathbf{Y}$ are not independent of each other (when constructed from an SCM, they both involve the same variable U), and so neither are $\mathrm{ICE}_{Z\to X}$ and $\mathrm{ICE}_{X\to Y}$. So we cannot just take expectations of all terms in Equation 19, and Equation 20 is typically not valid.

To make further progress, other assumptions must be imposed, particularly monotonicity:

$$X(Z=1) \geq X(Z=0). \qquad 21.$$

That is to say, we do not have any defiers, for which both $X(Z=0)=1$ (treatment would be taken when not assigned) and $X(Z=1)=0$ (treatment would not be taken when assigned).

Even monotonicity is not sufficient to allow estimation of $\mathrm{ACE}_{X\to Y}$. However, it does allow a new interpretation of the right-hand side of Equation 20, for it implies that the $\mathrm{ICE}_{Z\to X}$ of Equation 16 is either 1 or 0. Thus, from Equation 19,

$$\begin{aligned}\mathrm{ACE}_{Z\to Y} &= \mathrm{E}(\mathrm{ICE}_{X\to Y} \mid \mathrm{ICE}_{Z\to X} = 1) \times \Pr(\mathrm{ICE}_{Z\to X} = 1)\\ &= \mathrm{E}(\mathrm{ICE}_{X\to Y} \mid \mathrm{ICE}_{Z\to X} = 1) \times \mathrm{E}(\mathrm{ICE}_{Z\to X}).\end{aligned}$$

It follows that

$$\mathrm{ACE}_{Z\to Y}/\mathrm{ACE}_{Z\to X} = \mathrm{E}(\mathrm{ICE}_{X\to Y} \mid \mathrm{ICE}_{Z\to X} = 1). \qquad 22.$$

The right-hand side of Equation 22 is termed the local average treatment effect (LATE). Under monotonicity, LATE is estimable from the data because the left-hand side of Equation 22 is.

2.9.2. Critical comments.

1. Considerations similar to those in Section 2.8.1 suggest that it would typically be appropriate to regard potential outcomes, and so ICEs, as varying from one occasion to another, with only their distribution remaining constant.
2. In general, the monotonicity assumption is untestable, since (under the assumptions of 1 above) $X(Z = 1)$ and $X(Z = 0)$ cannot both be observed on the same occasion. However, it must hold in the case of an availability trial, as in Example 3, where necessarily $X(Z = 0) = 0$. Another extreme case where it can be inferred is in the presence of a variable W, a complete mediator between Z and X, so that $X(Z) = X\{W(Z)\}$, where we have empirical evidence that, with probability 1, $W(Z = 1) = 1$ and $X(W = 0) = 0$. If $X(Z = 0) = 1$, then we deduce $W(Z = 0) = 1 = W(Z = 1)$, and so $X(Z = 1) = X(Z = 0) = 1$, and we have no defiers.
3. LATE is an ACE in a subgroup of the population: those for whom both $X(Z = 0) = 0$ and $X(Z = 1) = 1$. These are termed compliers, since they would take the treatment if assigned to do so and not take it if not assigned (in an availability setting, they are those who would take the treatment if assigned to do so). However, it is impossible to tell who belongs to this subpopulation by knowing only what treatment was assigned and what treatment was taken (in an availability trial, an individual who was assigned treatment and took it must be a complier, but we still cannot tell the status of an individual who was not assigned treatment). Indeed, assuming as in comment 1 that $\mathbf{X}$ will vary from occasion to occasion, so too will the group of compliers, so the relevance of LATE in practice is debatable. Even its definition, relying as it does on a cross-world comparison of potential outcomes under both $Z = 1$ and $Z = 0$, can be criticized as essentially metaphysical and unscientific (Dawid & Didelez 2012).
4. In cases such as Example 4, where Z is not directly causal for X, the notation $X(Z = z)$ is meaningless and the above analysis cannot even get started.

3. CAUSES OF EFFECTS

3.1. Introduction

Let us consider again the initial attribution example: Juanita took Lipitor 80 mg daily for 3 years and developed diabetes. Was that because she took Lipitor?

One way of formulating this CoE question is through what the courts sometimes refer to as the "but for" test: Is it the case that, but for her having taken Lipitor, the diabetes would not have developed? This immediately plunges us into counterfactual considerations. We know that, in the actual world, the Lipitor was taken and diabetes developed, and we are asked to contrast this with the outcome in a counterfactual world, in which (counter to the known facts) the Lipitor had not been taken. The problem, of course, is that the counterfactual world is, by definition, unobservable, and even its existence—certainly its uniqueness—is questionable.

Even in deciding on the exact question, choices have to be made. Juanita took 80 mg of Lipitor daily for three years. Did she develop diabetes because she took the 80-mg dose (the only one for which the court accepted general causation) rather than 40 mg? Did diabetes develop because she took the drug for three years rather than two years? Each such choice conjures up a different counterfactual world for comparison with this one. We also have a choice in what we consider the observed response: Is it that she developed diabetes at some point, or that she developed diabetes within one year of stopping Lipitor? Detailed specification is obviously important in cases where the response is death; since death is certain, even in a counterfactual world, we can never say that an individual would never have died, but for some exposure.[5]

Under the but for criterion, causation is understood as the case that, in the appropriate counterfactual world, where Juanita did not take the Lipitor (in the same way that she in fact did), she did not develop diabetes (in the relevant time frame). This is appropriate when the response is all or nothing. We can also consider cases with a continuous response, such as time to death, but then it is not so clear what the focus of our attention should be. We might ask, for example, whether death occurs later in the relevant counterfactual world than it actually did in this world (Greenland 1999).

Even when our variables have been carefully specified and the relevant counterfactual question formulated, it remains unclear just how to conceive of and structure the counterfactual world of interest. Lewis (1973) develops an approach based on the closest possible world to this one, save only for the change to the exposure—but this only shifts the problem and does not solve it. There appears to be an unresolvable ambiguity about our counterfactual contrast.

Clearly there are deep philosophical problems, as well as technical specification issues, besetting any approach to formulating a CoE problem. In the following, we deal only with the case of binary exposure and outcome variables, denoted by X and Y, respectively, assuming the above specification problems have been addressed. But there will still remain some ambiguity about the relevant counterfactual world, which will be reflected in ambiguity about the answer to the CoE question.

In Section 3.2, we introduce two approaches to relating the actual and counterfactual worlds: SCMs and stochastic causal models (StCMs). The former is essentially deterministic, while the latter allows some stochastic elements. However, both make assumptions that might be regarded as overly strong, leading to misleadingly precise answers to the CoE question. In Section 3.3, we show how each of these models can be reformulated in terms of potential outcomes. In Section 3.4, we explain how, taking full account of real-world data on exposure and outcome, this approach can handle and quantify the remaining ambiguities by supplying an appropriate interval of ambiguity for the PC.

We can narrow the interval of ambiguity for an individual case by gaining a deeper understanding of the mechanisms and processes involved (Beyea & Greenland 1999), even when we cannot access the specific details of these for the individual case at hand. We develop this theme in the remaining sections, showing how information about additional variables can tighten the bounds on PC.

3.2. Counterfactual Constructions

Here we consider two formal approaches that support counterfactual assertions, the SCM, as already introduced, and the StCM.

3.2.1. Structural causal models. The approach of Pearl (2009) (see also Dawid et al. 2015, Pearl 2015) to CoE is based on SCMs. In the case of Juanita, this would involve the introduction

[5] Even in EoC cases, specification of the outcome matters. In March 2021, a few cases of blood clots among individuals who had received the AstraZeneca vaccine against COVID-19 were observed, and concerns were raised about a possible causal connection, leading to a pause in the vaccine's roll out in some countries. When it was pointed out that the rate of such clots was in fact lower than in the general population, attention turned to the few cases of a specific rare presentation of the blood clot, cerebral venous sinus thrombosis, and whether the vaccine could cause that.

of an unobserved exogenous background variable U, and the assumption that Juanita's diabetes status Y is fully determined by her Lipitor status X and U: $Y = f_Y(X, U)$ (this requires a conception of U as comprising all other preexisting quantities that, together with X, would totally determine Y—a collection that may not be easy to comprehend, let alone specify). In some contexts it might be appropriate to regard X as independent of U in the observational regime, the property of ignorability, as would happen if, for example, X is generated by a randomizing device. We do not impose this throughout and will specify where we do assume it.

The same functional relationship $Y = f_Y(X, U)$ is assumed to hold (consistency) whether or not X is imposed by external intervention. To this invariance requirement, familiar from EoC analysis, we add another, specific to CoE: that the value of the background variable U be the same in both the factual world and the counterfactual world that we wish to contrast with this one.

To start, we assume that the function f_Y and the joint distribution of (X, U) are known. These unrealistic requirements are removed in Section 3.4. In the factual world, we have observed $X = 1$ and $Y = 1$, i.e., $f_Y(1, U) = 1$. We can express the resulting uncertainty about the value of U by means of its conditional distribution, given $X = 1, f_Y(1, U) = 1$ [or, under ignorability, given only $f_Y(1, U) = 1$]. We now turn to consider the counterfactual world. Although U is supposed to be the same in both worlds (and thus endowed with the above conditional distribution), X and Y need not be. We introduce mirror variables X' and Y' as their counterfactual counterparts. We retain the general structure across worlds, so that $Y' = f_Y(X', U)$, both in observational and in interventional counterfactual regimes.

In the counterfactual world, we now consider the effect of an intervention $X' \leftarrow 0$. The value of Y' will be $f_Y(0, U)$. Using the previous conditional distribution of U, we obtain the counterfactual distribution for Y', given the hypothetical intervention $X' \leftarrow 0$ and the factual knowledge $X = Y = 1$. We can thus evaluate the PC as the probability that, in this distribution, $Y' = 0$; that is,

$$\mathrm{PC} = \Pr(Y' = 0 \mid X = 1, Y = 1, X' \leftarrow 0). \qquad 23.$$

3.2.2. Stochastic causal models.

A generalization of the above model, which does not require deterministic functional relationships, was suggested by Dawid (2000, section 12), although it has not been developed in detail. We again assume that the variable U retains its identity across the parallel worlds and introduce the mirror variables X' and Y'. But we now allow the dependence of Y on (X, U) to be given by a known conditional probability distribution; this allows a more liberal attitude to the nature of U, which can be a perfectly normal and in principle observable (though typically not observed) variable.

We assume that the same stochastic relationship also governs the dependence of Y' on (X', U). We again require consistency to relate observational and interventional regimes: The value of Y (respectively, Y') would be the same, whether X (respectively, X') arose by intervention or not. In order to complete the specification, we regard Y' and Y as conditionally independent, given (X, X', U) (for example, we might consider them as involving random noise, operating independently across worlds).

Having now a joint distribution for all variables in all worlds, we can again compute $\mathrm{PC} = \Pr(Y' = 0 \mid X = 1, Y = 1, X' \leftarrow 0)$. Specifically, by conditional independence,

$$\begin{aligned}\Pr(Y' = 0 \mid X = 1, Y = 1, X' \leftarrow 0, U) &= \Pr(Y' = 0 \mid X' \leftarrow 0, U)\\ &= \Pr(Y' = 0 \mid X' = 0, U)\\ &= \Pr(Y = 0 \mid X = 0, U).\end{aligned}$$

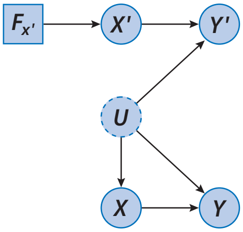

Figure 4

Twin network. Under ignorability, the arrow from U to X can be removed. The one from U to X' is absent because we are considering X' as set by external intervention, taking no account of U.

Then PC is the expectation of this, in the conditional distribution of U given $X = 1$, $Y = 1$, $X' \leftarrow 0$. But setting X' does not affect the joint distribution of (X, Y, U), so we just compute, from Bayes' theorem, the posterior density

$$p(u \mid X = 1, Y = 1) \propto \Pr(X = 1 \mid u)\ \Pr(Y = 1 \mid X = 1, u)\, p(u),$$

where $p(u)$ is the prior density of U (and the first term on the right-hand side can be omitted under ignorability). Finally,

$$\mathrm{PC} = \frac{\int \Pr(Y = 0 \mid X = 0, u)\ \Pr(X = 1 \mid u)\ \Pr(Y = 1 \mid X = 1, u)\, p(u)\, \mathrm{d}u}{\int \Pr(X = 1 \mid u)\ \Pr(Y = 1 \mid X = 1, u)\, p(u)\, \mathrm{d}u}. \qquad 24.$$

3.2.3. Twin network. In both the SCM and StCM approaches, the required computation can be automated by building a twin network representation of the problem (Pearl 2009, section 7.1.4), as in **Figure 4**, and making use of probability propagation algorithms (Cowell et al. 1999) as implemented in software systems such as `HUGIN` (**https://www.hugin.com**). This is useful in more complex problems with more variables.

The factual information $X = 1$ and $Y = 1$ and the counterfactual intervention $F_{X'} = 0$ (implying $X' = 0$) are entered at the relevant nodes and propagated through the network to obtain the appropriate conditional distribution for Y'.

3.2.4. More general StCMs. The StCM approach involves designating some of the variables in a problem as shared across worlds, while the others are allowed to differ. The associated twin network will have a single copy of the shared variables and mirror copies of the others, with the original DAG replicated and stitched together through the shared variables. As discussed by Dawid (2000), the choice of which variables are to be regarded as shared[6] is a matter of imagination rather than science and should relate to the specific problem of interest—there can be no context-free right answer. For example, there have been lawsuits by various states against tobacco companies claiming that if the companies had publicized their knowledge of the dangers of smoking when they first knew of them, many lives could have been saved. Damages have been sought for the additional cost placed on health services, meaning the excess cost in the actual world, over that of an imagined world in which the companies had made their knowledge public. But how should we imagine that world? One could reasonably argue that in such a world, as a result of giving up smoking, many people would have lived longer than they actually did, with consequential increased cost to health services. But what seems to be required for the case at hand is to imagine a world

[6] Mackie (1980) refers to these as the causal field and notes the role of our own choice in its specification.

where people had the same lifetimes but were healthier—i.e., to regard lifetimes as shared across parallel worlds—even though lifetimes can be considered an effect of the companies' decisions. It is not clear how such considerations could be accommodated in an SCM.

3.3. Potential Outcomes

As observed in Section 2.9, the SCM approach produces implied potential outcomes, $Y(1) = f_Y(1, U)$ and $Y(0) = f_Y(0, U)$. The pair $\mathbf{Y} = (Y(1), Y(0))$ is a function of U, with a bivariate distribution induced by that of U. And, in fact, $\mathbf{Y}$ is all that needs to be retained of U to fully describe the problem: We can replace U by $\mathbf{Y}$, with the functional dependence of Y on $(X, \mathbf{Y})$ given simply by $Y = Y(X)$. The problem can thus be more concisely expressed in terms of $(X, \mathbf{Y})$, with these having a joint distribution. Then Equation 23 becomes

$$\mathrm{PC} = \Pr(Y(0) = 0 \mid X = 1, Y(1) = 1). \qquad 25.$$

Note that under ignorability, X is independent of $\mathbf{Y}$ (this is indeed the very definition of ignorability in the PO framework), and then the conditioning on $X = 1$ in Equation 25 can be removed.

For the StCM approach we proceed as follows. The stochastic dependence of Y on (X, U) can be modeled by introducing a further unobserved noise variable V, independent of (X, U), and representing $Y = f_Y(X, U, V)$ for a suitable function f_Y. This can be done in many ways. One possible way uses the probability integral transformation: If Y is a univariate variable whose conditional distribution function $F_{x,u}(y)$, given $X = x$, $U = u$, is strictly increasing, take V to be uniform on $[0, 1]$, and $f_Y(x, u, v) = F_{x,u}^{-1}(v)$.

There will be a counterfactual mirror V' of V, with $Y' = f_Y(X, U, V')$. We now define potential outcomes $Y(1) = f_Y(1, U, V)$ and $Y(0) = f_Y(0, U, V')$, having a joint distribution induced by that of (U, V, V'). Although the variables so constructed will depend on the specific choices made for the noise variable V and the function f_Y, it is easy to see that their joint distribution will in all cases be that of (Y, Y'), given interventions $X \leftarrow 1, X' \leftarrow 0$. And since $X = 1 \Rightarrow Y = Y(1)$, etc., Equation 25 again holds.

We thus see that, in all cases, we can ignore the finer details and represent the problem by means of a joint distribution for $(X, \mathbf{Y})$, with PC given by Equation 25.

3.4. Empirical Information

We have so far supposed that the full probabilistic structure of the model, with its variables U, X, Y, is known. In an StCM, we can take U to be a specified potentially observable variable, and then this assumption is not unreasonable. However, it is typically implausible for an SCM, where, in order to achieve the required deterministic dependence of Y on (X, U), we would have to conceive of a fantastically rich U. Alternatively, we can reexpress the problem in terms of the pair $\mathbf{Y}$ of potential responses, with a joint distribution for $(X, \mathbf{Y})$. Without making further assumptions on the originating SCM or StCM, there are no constraints on this joint distribution (other than independence of X and $\mathbf{Y}$ under ignorability). We can, however, gather empirical data to constrain it and thus hope to estimate PC by Equation 25.

In the following, we proceed on this basis and consider what can indeed be estimated. We initially assume that we can only observe X and Y, in interventional and/or observational circumstances. We can estimate $\Pr\{Y(x) = 1\} = \Pr(Y = 1 \mid X \leftarrow x)$ from interventional studies. When we cannot assume ignorability, we can also estimate, from observational data, $\Pr\{Y(1) = 1 \mid X = 1\} = \Pr(Y = 1 \mid X = 1)$ (by consistency) and $\Pr\{Y(0) = 1 \mid X = 0\} = \Pr(Y = 1 \mid X = 0)$, as well as the marginal distribution of X. For this general case it might initially seem problematic to estimate, say, $\Pr\{Y(1) = 1 \mid X = 0\}$, since this involves noncoexisting worlds, one with $X = 1$ and the other with $X = 0$; but we can, in fact, solve for it, using $\Pr\{Y(1) = 1\} = \Pr\{Y(1) = 1 \mid X = 1\} \Pr(X = 1) + \Pr\{Y(1) = 1 \mid X = 0\} \Pr(X = 0)$, where all other terms are estimable.

We can thus estimate the bivariate distribution of $(X, Y(1))$ and, likewise, that of $(X, Y(0))$. However, the full trivariate distribution of $(X, Y(1), Y(0))$ is not estimable: Since we can never observe both $Y(0)$ and $Y(1)$ simultaneously, no data can tell us directly about the dependence between $Y(0)$ and $Y(1)$, either marginally or conditionally on X. And since Equation 25 requires such information, typically PC is not identifiable from empirical data.

Nevertheless, the estimable bivariate distributions do impose constraints on the possible values of PC. Moreover, these constraints can often be tightened still further when we can observe other, related variables in the problem. We now turn to investigating such constraints in a number of contexts.

3.5. Analysis of the Probability of Causation

Let us consider the initial attribution example: Juanita took 80 mg of Lipitor daily for three years ($X = 1$) and developed diabetes ($Y = 1$). Was that because she took Lipitor? We wish to address this question and assess the PC for Juanita's case, using data collected on other individuals. To this end, we assume exchangeability: Juanita is similar to the population from which probabilities have been computed, so that those probabilities apply to her. Exchangeability may require restriction of the data considered to individuals deemed sufficiently like Juanita.

Except where the assumption is relaxed in Section 3.5.3, we also assume ignorability: The fact that Juanita chose to take the drug is not informative about her response to it, either factually or counterfactually. Formally, we require independence, $X \perp\!\!\!\perp \mathbf{Y}$, between X and the pair of potential responses $\mathbf{Y} = (Y(1), Y(0))$. Ignorability is a strong assumption and will often be inappropriate. When it can be assumed, we can use data from either experimental or observational studies; otherwise, we need data from both of these.

Under ignorability, the target Equation 25 becomes

$$\mathrm{PC} = \Pr(Y(0) = 0 \mid Y(1) = 1). \qquad 26.$$

We proceed to assess this using the general potential outcome framework of Sections 3.3 and 3.4, where no assumptions are imposed on the joint distribution of $Y(0)$ and $Y(1)$ beyond those that can be informed by the empirical data. Further details are provided by Dawid et al. (2017) and Dawid & Musio (2022).

3.5.1. Basic inequalities.

Suppose we have access to (observational or experimental) data, supplying values for

$$\Pr\{Y(x) = y\} = \Pr(Y = y \mid X = x) \quad (x = 0, 1 \quad y = 0, 1). \qquad 27.$$

Define

$$\tau := \Pr(Y = 1 \mid X = 1) - \Pr(Y = 1 \mid X = 0),$$
$$\rho := \Pr(Y = 1 \mid X = 1) - \Pr(Y = 0 \mid X = 0).$$

The joint distribution of $(Y(0), Y(1))$ must have the form of **Table 1**, where the marginal probabilities are given by Equation 27, reexpressed in terms of τ and ρ, and where the unidentified slack quantity ξ embodies the residual ambiguity in the distribution. For all the entries of **Table 1**

Table 1 Joint distribution of $Y(0)$ and $Y(1)$

	$Y(1) = 0$	$Y(1) = 1$	
$Y(0) = 0$	$\frac{1}{2}(1 - \rho - \xi)$	$\frac{1}{2}(\xi + \tau)$	$\frac{1}{2}(1 + \tau - \rho)$
$Y(0) = 1$	$\frac{1}{2}(\xi - \tau)$	$\frac{1}{2}(1 + \rho - \xi)$	$\frac{1}{2}(1 - \tau + \rho)$
	$\frac{1}{2}(1 - \tau - \rho)$	$\frac{1}{2}(1 + \tau + \rho)$	1

to be nonnegative, we require

$$|\tau| \leq \xi \leq 1 - |\rho|. \tag{28.}$$

The PC (Equation 26) is

$$\mathrm{PC} = \frac{\xi + \tau}{1 + \tau + \rho}. \tag{29.}$$

On using Inequalities 28, we obtain the following interval bounds for PC:

$$l := \max\left\{0, \frac{2\tau}{1 + \tau + \rho}\right\} \leq \mathrm{PC} \leq \min\left\{1, \frac{1 + \tau - \rho}{1 + \tau + \rho}\right\} =: u, \tag{30.}$$

or equivalently,

$$l = \max\left\{0, 1 - \frac{1}{\mathrm{RR}}\right\} \leq \mathrm{PC} \leq \min\left\{1, \frac{\Pr(Y = 0 \mid X = 0)}{\Pr(Y = 1 \mid X = 1)}\right\} = u, \tag{31.}$$

where the risk ratio (RR) (Robins & Greenland 1989) is

$$\mathrm{RR} = \frac{\Pr(Y = 1 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)}. \tag{32.}$$

In the absence of additional information or assumptions, these bounds constitute the best available inference regarding PC. In particular, RR > 2, doubling the risk, implies that PC > 0.5. In a civil legal case, causality might then be concluded on the balance of probabilities. However, because of the remaining ambiguity, expressed by Inequalities 31, finding that RR falls short of 2 does not imply that PC < 0.5.

The above assumes that our data are sufficiently extensive to support precise estimation of the probabilities appearing in the bounds of Inequalities 31, and even then we have residual ambiguity. With more limited data this ambiguity is compounded with statistical uncertainty. A Bayesian approach to inference in such cases is presented by Dawid et al. (2016b).

3.5.2. Refining the inequalities: covariates.

When we have additional information, we may be able to refine our inferences about PC (Kuroki & Cai 2011).

Thus, suppose that we also have information on a covariate S, a pretreatment individual characteristic that can vary from person to person and can have an effect on both X and Y. For simplicity we suppose that S is discrete and that we can estimate from the data the full joint distribution of (S, X, Y). We assume exchangeability, and ignorability conditional on S. The introduction of such information involves consideration of what is an appropriate reference population: Among all the possible groups of individuals, with various specifications for the covariates, find that best fitting the conditions of exchangeability and ignorability. This problem can be addressed by a Bayesian model selection procedure (Corradi & Musio 2020).

The relevant potential responses are now $\mathbf{X} := (X(s) : s \in S)$ and $\mathbf{Y} := (Y(s, x) : s \in S, x = 0 \text{ or } 1)$, and the relationship between potential and actual responses is $X = X(S)$ and $Y = Y(S, X)$. Ignorability is now formalized as requiring mutual independence between S, $\mathbf{X}$, and $\mathbf{Y}$.

In the case that we are able to measure S for Juanita, say $S = s$, we can simply restrict the experimental subjects to those having the same covariate value (who are thus like Juanita in all relevant respects). The PC is now

$$\mathrm{PC}(s) = \Pr\left(Y(s, 0) = 0 \mid Y(s, 1) = 1\right).$$

We can bound this just as in Inequalities 31, but with all probabilities now conditioned on $S = s$, obtaining $l(s) \leq \mathrm{PC}(s) \leq u(s)$.

More interesting is the case in which we do not observe S for Juanita. We have to consider what would have been the response if, counterfactually, Juanita's exposure had been $X = 0$. We assume that this is the minimal change made between the factual and the counterfactual worlds so that, in particular, there is no change to the value or distribution of S.

The PC is now

$$\begin{aligned}\mathrm{PC} &= \Pr\left\{Y(S, 0) = 0 \mid X(S) = 1, Y(S, 1) = 1\right\}\\ &= \sum_s \mathrm{PC}(s) \times \Pr(S = s \mid X = 1, Y = 1).\end{aligned} \qquad 33.$$

There are no logical relationships between the distributions of $(Y(s, 0), Y(s, 1))$ for different values of S. So by independently varying the values taken by the slack variables in the joint distribution of these potential responses, all the lower bounds $l(s)$ for $\mathrm{PC}(s)$ can be achieved simultaneously. This leads to an achievable lower bound for PC:

$$\mathrm{PC} \geq L = \sum_s l(s) \times \Pr(S = s \mid X = 1, Y = 1). \qquad 34.$$

We can express

$$\begin{aligned}L = &\frac{1}{\Pr(Y = 1 \mid X = 1)}\\ &\times \sum_s \max\left\{0, \Pr(Y = 1 \mid X = 1, S = s) - \Pr(Y = 1 \mid X = 0, S = s)\right\} \times \Pr(S = s \mid X = 1).\end{aligned} \qquad 35.$$

Similarly, we obtain upper bound

$$\begin{aligned}U = 1 - &\frac{1}{\Pr(R = 1 \mid E = 1)}\\ &\times \sum_s \max\{0, \Pr(Y = 1 \mid X = 1, S = s) - \Pr(Y = 0 \mid X = 0, S = s)\} \times \Pr(S = s \mid X = 1).\end{aligned} \qquad 36.$$

We cannot compare these bounds directly with those of Inequalities 31 since, when we do not take account of S, the relation between X and Y is generally nonignorable: $\Pr(Y = y \mid X \leftarrow x) \neq \Pr(Y = y \mid X = x)$.

3.5.3. Nonignorability.

Tian & Pearl (2000) analyze the nonignorable case where we do not observe S, either for Juanita or in the external data (still considered exchangeable with Juanita). We now need both observational and experimental data on X and Y. Tian & Pearl (2000) develop the following lower bound for $\mathrm{PC} = \Pr\{Y(0) = 0 \mid X = 1, Y(1) = 1\}$:

$$L' = \max\left\{0, \frac{\Pr(Y = 1) - \Pr(Y = 1 \mid X \leftarrow 0)}{\Pr(X = 1, Y = 1)}\right\}. \qquad 37.$$

Dawid & Musio (2022) show that this can also be derived as a special case of our Equation 35 if we substitute for S the binary variable D, desired exposure (Corradi & Musio 2020). D will be identical to X in an observational context but need not be so in an experimental setting, where D may not be observable.

In our case, with access to information on S, we could compute $\Pr(Y = 1 \mid X \leftarrow 0)$ by the back-door formula,

$$\Pr(Y = 1 \mid X \leftarrow 0) = \sum_{s} \Pr(Y = 1 \mid X = 0, S = s) \times \Pr(S = s), \tag{38.}$$

and thus compute L' of Equation 37. It can be shown (Dawid & Musio 2022) that $L' \leq L$ of Equation 35, with equality if and only if all the conditional RRs

$$\frac{\Pr(Y = 1 \mid X = 1, S = s)}{\Pr(Y = 1 \mid X = 0, S = s)}, \qquad (s \in S),$$

lie on the same side of 1: Knowing, and using, the information about S is at least as good as ignoring it. Similarly, we can show that the upper bound U of Equation 36 does not exceed the upper bound U' derived by Tian & Pearl (2000),

$$U' = \min\left\{1, \frac{\Pr(Y = 0 \mid X \leftarrow 0) - \Pr(X = 0, Y = 0)}{\Pr(X = 1, Y = 1)}\right\},$$

with equality if and only if all the ratios

$$\frac{\Pr(Y = 1 \mid X = 1, S = s)}{\Pr(Y = 0 \mid X = 0, S = s)}, \qquad (s \in S),$$

lie on the same side of 1.

3.5.4. Mediators. We now consider the case that a third variable M acts as a complete mediator in the causal pathway $X \rightarrow M \rightarrow Y$ between the exposure X and the response Y. Again we restrict to the case that all variables are binary. We introduce the potential value $M(x)$ of M for $X = x$, and $Y(m)$, the potential value of Y for $M = m$, and define $\mathbf{M} = (M(0), M(1))$ and $\mathbf{Y} = (Y(0), Y(1))$. We observe X, $M = M(X)$, and $Y = Y(M)$. We assume the exchangeability and the ignorability conditions, the latter expressed as mutual independence between X, $\mathbf{M}$, and $\mathbf{Y}$. This implies the observational conditional independence

$$Y \perp\!\!\!\perp X \mid M, \tag{39.}$$

which is a testable implication of our assumptions. We assume we have data supplying values for $\Pr(M = m \mid X = x)$ and $\Pr(Y = y \mid M = m)$ and compute, by Property 39,

$$\Pr(Y = y \mid X = x) = \sum_{m} \Pr(Y = y \mid M = m) \Pr(M = m \mid X = x). \tag{40.}$$

For the case that M is observed in the experimental data but not for Juanita, Dawid et al. (2016a) showed that this additional information does not change the lower bound l on PC in Inequalities 31 but does lower the upper bound u. Dawid & Musio (2022) extend this analysis to cases with additional covariates, while Dawid et al. (2022) deal with the case that we have a complete mediation sequence $X = M_0 \rightarrow M_1 \rightarrow \cdots \rightarrow M_{n-1} \rightarrow M_n = Y$, and know the probabilistic structure of each link in the chain, with all, some, or none of the *M*s being observed for Juanita.

3.6. Further Causes of Effects Problems

We have only dealt here with the case of a single putative cause, understanding causation in terms of the but for criterion. There are many more complex problems that cannot be handled in this way: In particular, the whole field of legal causation has to handle a wide variety of problems involving multiple competing causes and other concepts of causality (Hart & Honoré 1985, Goldberg 2011). While there have been some interesting statistical treatments of specific problems (e.g., Cox 1984), it seems fair to say that general philosophical understandings of causality in such problems have not reached maturity. To the extent that problems are modeled in formal terms, this often involves a purely deterministic understanding of causality, which is not easily translated into a stochastic framework. Halpern (2016) makes an interesting attempt to pin down the concept of the actual cause using the SCM framework but admits that he is unable to reach a fully satisfying conclusion.

There is clearly much ground remaining to be covered in understanding CoE, but it is perhaps premature to attempt more detailed statistical treatment before clearer general principles have emerged.

4. CONCLUSION

4.1. What Have We Achieved?

We have taken a tour through a variety of formalisms by means of which problems of statistical causality are currently conceived and modeled. We have considered decision theory, linear SEMs, SCMs, StCMs, and POs, together with their associated graphical representations. We have paid particular attention to the distinction between inference about EoC and inference about CoE, and the pros and cons of applying the various formalisms to address each of these tasks.

Most current treatments of statistical causality use identical frameworks (for example, PO and SCM) to address both EoC and CoE problems. We argue that this is mistaken. It is important to consider carefully which tools are fit for which tasks. The use of an inappropriate formal framework brings with it the danger that we uncritically treat any mathematically well-formed formula (such as those describing ICE or LATE) as meaningful, when it need not be.

EoC analysis has no need for counterfactual reasoning and thus no need for any of the formalisms, such as POs and SCMs, which can support this: All that is required is stochastic modeling and statistical decision theory, as in the DT approach. Other approaches combine philosophically questionable ingredients, unnecessarily complex arguments, and the risk of making misleading inferences.

Some form of counterfactual framework appears unavoidable for representing and analyzing CoE. All the approaches considered, other than the DT approach, support such reasoning by allowing formal statements concerning two or more parallel worlds simultaneously. But the deterministic world view implicit in the SCM approach is inessential, as is evidenced by the alternative StCM formalism.

Whether we start from an SCM or StCM, we can in each case construct a derived PO representation. This can then be used to characterize the essential ambiguities in identifying a PC from data—such ambiguities depend both on how the CoE question is conceived and formulated, and on what variables are observed in the data. We have shown how the degree of ambiguity can be reduced by detailed knowledge of mechanisms, in particular probabilistic relationships between cause, effect, and additional variables such as covariates or mediators, even when those additional variables are not measured for the individual case at hand.

4.2. Lipitor and Diabetes

We close with brief remarks on the issues of general and specific causation as presented in the legal judgement on Lipitor (*Hempstead v. Pfizer, Inc.* 2015).

- General causation: "… Epidemiologists use a two-step process for establishing general causation… First, studies must establish an association or correlation between two variables, here, Lipitor and diabetes… Once an association is established, epidemiologists apply the 'Hill factors' to evaluate whether an association is causal" (*Hempstead v. Pfizer, Inc.* 2015). This shows some of the limitations of our own analysis. We have assumed that available data are sufficiently extensive that identifiable probabilities can be estimated essentially perfectly. In the real world, with statistical uncertainty arising from having only finite data, other questions arise, such as whether an association, causal or not, exists at all. Then, "to evaluate whether an association is causal," we would ideally conduct experiments or argue from observational data by means of defensible assumptions. But this may not be possible, and then more informal, pragmatic approaches, such as those embodied in the Hill factors (Hill 1965), can be invoked to provide some support for causation.
- Specific causation: "… Even if Plaintiffs establish that there is an association between Lipitor and diabetes… it does not necessarily follow the Lipitor caused the development of diabetes in a particular plaintiff… In epidemiological terms, a two-fold increased risk is an important showing for plaintiffs to make because it is the equivalent of the required legal burden of proof—a showing of causation by the preponderance of the evidence or, in other words, a probability of greater than 50%" (*Hempstead v. Pfizer, Inc.* 2015). Here we see an application of the lower bound of Inequalities 31. In the case of Juanita, the relevant relative risk was estimated as 1.6. This was regarded as establishing general causation; however, as far as specific causation is concerned, it implies only that the PC for Juanita is at least 0.38 (the upper bound being 1) and thus does not logically rule out that its value could still be greater than 50%. However, from a legal, rather than logical, standpoint it does not seem unreasonable to require that the lower bound exceed 50%.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

M.M. was partially supported by the project STAGE (Statistics Algebra and Geometry) of Fondazione di Sardegna.

LITERATURE CITED

Angrist J, Imbens G, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Balke AA, Pearl J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92:1172–76

Beyea J, Greenland S. 1999. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Phys.* 76:269–74

Bowden RJ, Turkington DA. 1984. *Instrumental Variables*. Cambridge, UK: Cambridge Univ. Press

Bühlmann P. 2020. Invariance, causality and robustness. *Stat. Sci.* 35:404–26

Constantinou P, Dawid AP. 2017. Extended conditional independence and applications in causal inference. *Ann. Stat.* 45:2618–53

Corradi F, Musio M. 2020. Causes of effects via a Bayesian model selection procedure. *J. R. Stat. Soc. Ser. A* 183:1777–92

Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. 1999. *Probabilistic Networks and Expert Systems*. New York: Springer

Cox LA. 1984. Probability of causation and the attributable proportion of risk. *Risk Anal.* 4:221–30

Cuellar M. 2017. *Causal reasoning and data analysis in the law: definition, estimation, and usage of the probability of causation*. PhD Thesis, Dep. Stat. and Heinz Coll. Public Policy, Carnegie Mellon Univ., Pittsburgh, PA

Dawid AP. 1976. Properties of diagnostic data distributions. *Biometrics* 32:647–58

Dawid AP. 1979. Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. B* 41:1–31

Dawid AP. 2000. Causal inference without counterfactuals (with discussion). *J. Am. Stat. Assoc.* 95:407–48

Dawid AP. 2003. Causal inference using influence diagrams: the problem of partial compliance (with discussion). In *Highly Structured Stochastic Systems*, ed. PJ Green, NL Hjort, S Richardson, pp. 45–81. Oxford, UK: Oxford Univ. Press

Dawid AP. 2010. Beware of the DAG! *J. Mach. Learn. Res.* 6:59–86

Dawid AP. 2013. Vote of thanks on "A Bayesian approach to complex clinical diagnoses: a case-study in child abuse," by Nicky Best, Deborah Ashby, Frank Dunstan, David Foreman and Neil McIntosh. *J. R. Stat. Soc. Ser. A* 176:83–84

Dawid AP. 2015. Statistical causality from a decision-theoretic perspective. *Annu. Rev. Stat. Appl.* 2:273–303

Dawid AP. 2017. On individual risk. *Synthese* 194:3445–74

Dawid AP. 2021. Decision-theoretic foundations for statistical causality. *J. Causal Inference* 9:39–77

Dawid AP. 2022. The tale wags the DAG. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, ed. R Dechter, H Geffner, J Halpern. New York: ACM. In press

Dawid AP, Didelez V. 2012. "Imagine a can opener"—the magic of principal stratum analysis. *Int. J. Biostat.* 8(1):19

Dawid AP, Faigman DL, Fienberg SE. 2014. Fitting science into legal contexts: assessing effects of causes or causes of effects? (with discussion and authors' rejoinder). *Sociol. Methods Res.* 43:359–421

Dawid AP, Faigman DL, Fienberg SE. 2015. On the causes of effects: response to Pearl. *Sociol. Methods Res.* 44:165–74

Dawid AP, Humphreys M, Musio M. 2022. Bounding causes of effects with mediators. *Sociol. Methods Res.* In press

Dawid AP, Murtas R, Musio M. 2016a. Bounding the probability of causation in mediation analysis. In *Topics on Methodological and Applied Statistical Inference*, ed. TD Battista, E Moreno, W Racugno, pp. 75–84. New York: Springer

Dawid AP, Musio M. 2022. What can group level data tell us about individual causality? In *Statistics in the Public Interest*, ed. A Carriquiry, J Tanur, W Eddy. New York: Springer. In press

Dawid AP, Musio M, Fienberg SE. 2016b. From statistical evidence to evidence of causality. *Bayesian Anal.* 11:725–52

Dawid AP, Musio M, Murtas R. 2017. The probability of causation. *Law Probab. Risk* 16:163–79

Faigman DL, Monahan J, Slobogin C. 2014. Group to individual (G2i) inference in scientific expert testimony. *Univ. Chicago Law Rev.* 81:417–80

Goldberg R, ed. 2011. *Perspectives on Causation*. Oxford, UK: Hart

Greenland S. 1999. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am. J. Public Health* 89:1166–69

Halpern JY. 2016. *Actual Causality*. Cambridge, MA: MIT Press

Hart HLA, Honoré AM. 1985. *Causation in the Law*. Oxford, UK: Clarendon

Hausman D. 1998. *Causal Asymmetries*. Cambridge, UK: Cambridge Univ. Press

Hempstead v. Pfizer, Inc., 150 F. Suppl. 3d 644 (D.S.C. 2015)

Hill AB. 1965. The environment and disease: association or causation? *Proc. R. Soc. Med.* 58:295–300

Holland PW. 1986. Statistics and causal inference (with discussion). *J. Am. Stat. Assoc.* 81:945–70

Holland PW. 1988. Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.* 18:449–84

Honoré AM. 2010. Causation in the law. In *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*, ed. EN Zalta. Stanford, CA: Stanford Univ. Metaphys. Res. Lab. **https://plato.stanford.edu/archives/win2010/entries/causation-law/**

Imbens GW, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–76

Katan MB. 1986. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 327(8479):507–8

Kuroki M, Cai Z. 2011. Statistical analysis of 'probabilities of causation' using co-variate information. *Scand. J. Stat.* 38:564–77

Lauritzen SL, Dawid AP, Larsen BN, Leimer HG. 1990. Independence properties of directed Markov fields. *Networks* 20:491–505

Lewis DK. 1973. *Counterfactuals*. Oxford, UK: Blackwell

Mackie JL. 1980. *The Cement of the Universe: A Study of Causation*. Oxford, UK: Oxford Univ. Press

Mill JS. 1843. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and Methods of Scientific Investigation*. London: John W. Harper

Pearl J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.

Pearl J. 2015. Causes of effects and effects of causes. *Sociol. Methods Res.* 44:149–64

Pearl J, Mackenzie D. 2018. *The Book of Why*. New York: Basic Books

Price H. 1991. Agency and probabilistic causality. *Br. J. Philos. Sci.* 42:157–76

Reichenbach H. 1956. *The Direction of Time*. Berkeley: Univ. Calif. Press

Robins JM, Greenland S. 1989. The probability of causation under a stochastic model for individual risk. *Biometrics* 45:1125–38

Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701

Sanders J, Faigman DL, Imrey PB, Dawid AP. 2021. Differential etiology: inferring specific causation in the law from group data in science. *Arizona Law Rev.* 63:851–922

Tian J, Pearl J. 2000. Probabilities of causation: bounds and identification. 28:287–313

Verma T, Pearl J. 1990. Causal networks: Semantics and expressiveness. In *Machine Intelligence and Pattern Recognition*, Vol. 9, ed. RD Shachter, TS Levitt, LN Kanal, JF Lemmer, pp. 69–76. Amsterdam: Elsevier

Woodward J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford, UK: Oxford Univ. Press

Woodward J. 2016. Causation and manipulability. In *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Stanford, CA: Stanford Univ. Metaphys. Res. Lab. **https://plato.stanford.edu/entries/causation-mani/**

Wright SS. 1921. Correlation and causation. *J. Agric. Res.* 20:557–85