Causal diagrams for empirical research
(with appendix on structural equations and counterfactuals)

JUDEA PEARL
Cognitive Systems Laboratory, Computer Science Department,
University of California, Los Angeles, CA 90024, U.S.A.

SUMMARY

The primary aim of this paper is to show how graphical models can be used as a mathematical language for integrating statistical and subject-matter information. In particular, the paper develops a principled, nonparametric framework for causal inference, in which diagrams are queried to determine if the assumptions available are sufficient for identifying causal effects from nonexperimental data. If so the diagrams can be queried to produce mathematical expressions for causal effects in terms of observed distributions; otherwise, the diagrams can be queried to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.
Key words: Causal inference, graph models, structural equations, treatment effect.

1 INTRODUCTION

The tools introduced in this paper are aimed at helping researchers communicate qualitative assumptions about cause-effect relationships, elucidate the ramifications of such assumptions, and derive causal inferences from a combination of assumptions, experiments, and data.

The basic philosophy of the proposed method can best be illustrated through the classical example due to Cochran (Wainer, 1989). Consider an experiment in which soil fumigants (X) are used to increase oat crop yields (Y) by controlling the eelworm population (Z) but may also have direct effects (both beneficial and adverse) on yields beside the control of eelworms. We wish to assess the total effect of the fumigants on yields when this typical study is complicated by several factors. First, controlled randomized experiments are infeasible – farmers insist on deciding for themselves which plots are to be fumigated. Second, farmers' choice of treatment depends on last year's eelworm population (Z_0), an unknown quantity which is strongly correlated with this year's population — thus we have a classical case of confounding bias, which interferes with the assessment of treatment effects, regardless of sample size. Fortunately, through laboratory analysis of soil samples, we can determine the eelworm populations before and after the treatment and, furthermore, because the fumigants are known to be active for a short period only, we can safely assume that they do not affect the growth of eelworms surviving the treatment. Insead, eelworms' growth depends on the population of birds (and other predators) which is correlated, in turn, with last year's eelworm population and hence with the treatment itself.

The method proposed in this paper permits the investigator to translate complex considerations of this sort into a formal language, thus facilitating the following tasks:

1. Explicate the assumptions underlying the model.
2. Decide whether the assumptions are sufficient for obtaining consistent estimates of the target quantity: the total effect of the fumigants on yields.
3. If the answer to item (2) is affirmative, the method provides a closed-form expression for the target quantity, in terms of distributions of observed quantities.
4. If the answer to item (2) is negative, the method suggests a set of observations and experiments which, if performed, would render a consistent estimate feasible.

The first step in this analysis is to construct a causal diagram such as the one given in Figure 1 which represents the investigator's understanding of the major causal

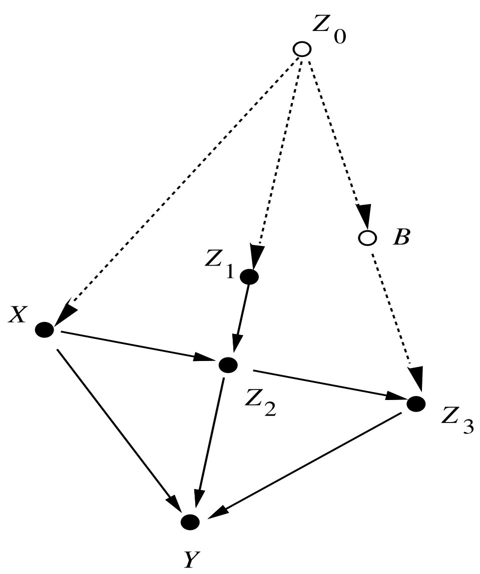

Figure 1: A causal diagram representing the effect of fumigants (X) on yields (Y).

influences among measurable quantities in the domain. For example, the quantities Z_1, Z_2, and Z_3 represent, respectively, the eelworm population (both size and type) before treatment, after treatment, and at the end of the season. Z_0 represents last year's eelworm population; because it is an unknown quantity, it is denoted by a hollow circle, as is the quantity B, the population of birds and other predators. Links in the diagram are of two kinds: those that connect unmeasured quantities are designated by dashed arrows, those connecting measured quantities by solid arrows. The substantive assumptions embodied in the diagram are negative causal assertions which are conveyed through the links missing from the diagram. For example, the missing arrow between Z_1 and Y signifies the investigator's understanding that pre-treatment eelworms can not affect oat plats directly; their entire influence on oat yields is mediated by post-treatment conditions, namely Z_2 and Z_3. The purpose of the paper is not to validate or repudiate such domain-specific assumptions but, rather, to test whether a given set of assumptions is sufficient for quantifying causal effects from nonexperimental data, for example, estimating the total effect of fumigants on yields.

The causal diagram in Figure 1 is similar in many respects to the path diagrams devised by Wright (1921): both reflect the investigator's subjective and qualitative knowledge of causal influences in the domain, both employ directed acyclic graphs, and both allow for the incorporation of latent or unmeasured quantities. The major differences lie in the method of analysis. First, whereas path diagrams have been analyzed mostly in the context of additive linear models, causal diagrams permit arbitrary nonlinear interactions. In fact, the analysis of causal effects will be entirely nonparametric, entailing no commitment to a particular functional form for equations and distributions. Second, causal diagrams will be used not only as a passive language to specify assumptions but also as an active computational device through which the desired quantities will be derived. For example, the proposed method allows an investigator to inspect the diagram of Figure 1 and conclude immediately that:

1. The total effect of X on Y can be estimated consistently from the observed distribution of X, Z_1, Z_2, Z_3, and Y.

2. The total effect of X on Y (assuming discrete variables throughout) is given by the formula[1]

$$P(y|\hat{x}) = \sum_{z_1}\sum_{z_2}\sum_{z_3} P(y|z_2,\ z_3,\ x)P(z_2|z_1,\ x)\sum_{x'} P(z_3|z_1,\ z_2,\ x')P(z_1,\ x') \quad (1)$$

where $P(y|\hat{x})$ stands for the probability of achieving a yield level of $Y = y$ given that the treatment is set to level $X = x$ by external intervention.

3. A consistent estimation of the total effect of X on Y would not be feasible if Y were confounded with Z_3; however, confounding Z_2 and Y will not invalidate the formula for $P(y|\hat{x})$.

These conclusions can be obtained either by analyzing the graphical properties of the diagram or by performing a sequence of symbolic derivations, governed by the diagram, which gives rise to causal effect formulas such as Eq. (1).

The formal semantics of the causal diagrams used in this paper will be defined in Section 2, following review of directed acyclic graphs (DAGs) as a language for communicating conditional independence assumptions (Subsection 2.1). Subsection 2.2 introduces a causal interpretation of DAGs based on nonparametric structural

[1]The reader need not be intimidated if, at this point, the formula appears unfamiliar. After reading Section 4, the reader should be able to derive such formulas with greater ease than solving a pair of algebraic equations. Note that x' is merely an index of summation that ranges over the values of X.

equations and demonstrates their use in predicting the effect of interventions. An alternative formulation is then described where interventions are treated as variables in an augmented probability space (shaped by the causal diagram) from which causal effects are obtained by ordinary conditioning. Using either interpretation, it is possible to quantify how probability distributions will change as a result of external interventions and to identify conditions under which randomized experiments are not necessary. Section 3 will demonstrate the use of causal diagrams to control confounding bias in observational studies. We will establish two graphical conditions ensuring that causal effects can be estimated consistently from nonexperimental data. The first condition, named the back-door criterion, is equivalent to the ignorability condition of Rosenbaum & Rubin (1983). The second condition, named the front-door criterion, involves covariates that are affected by the treatment, and thus introduces new opportunities for causal inference. In Section 4, we introduce a symbolic calculus that permits the stepwise derivation of causal effect formulas of the type shown in Eq. (1). Using this calculus, Section 5 characterizes the class of graphs that permit the quantification of causal effects from nonexperimental data or from surrogate experimental designs.

2 GRAPHICAL MODELS AND THE MANIPULATIVE ACCOUNT OF CAUSATION

2.1 Graphs and conditional independence

The usefulness of directed acyclic graphs (DAGs) as economical schemes for representing conditional independence assumptions is well acknowledged in the literature (Pearl, 1988; Whittaker, 1990). This usefulness stems from the existence of graphical methods for identifying the conditional independence relationships that are implied by recursive product decompositions

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{2}$$

where $\mathbf{pa}_i$ stands for the realization of any set of variables that precede X_i in some order $(X_1, X_2, \ldots, X_n)$. If we construct a DAG in which the variables corresponding to $\mathbf{pa}_i$ are represented as the *parents* of X_i (also called adjacent predecessors or direct influences of X_i) then the independencies implied by the decomposition (2) can be read off the DAG using the following test:

DEFINITION 1. (d-separation) Let X, Y, and Z be three disjoint subsets of nodes in a DAG G, and let p be any path between a node in X and a node in Y. (By a path we mean any succession of arcs, regardless of their directions.) Z is said to block p if there is a node w on p satisfying one of the following two conditions:

(i) w has converging arrows (along p) and neither w nor any of its descendants are in Z, or,

(ii) w does not have converging arrows (along p) and w is in Z. Z is said to d-separate X from Y, in G, denoted $(X \perp\!\!\!\perp Y|Z)_G$, iff Z blocks every path from a node in X to a node in Y. □

It can be shown that there is a one-to-one correspondence between the set of conditional independencies, $X \perp\!\!\!\perp Y|Z$ (Dawid, 1979), implied by the recursive decomposition of Eq. (2) and the set of triples (X, Z, Y) that satisfy the d-separation criterion in G (Geiger et al., 1990). For example, the DAG of Figure 1 represents the decomposition

$$P(z_0, x, z_1, b, z_2, z_3, y) = P(z_0)P(x|z_0)P(z_1|z_0)P(b|z_0)P(z_2|x, z_1)P(z_3|z_2, b)P(y|x, z_2, z_3)$$

and it implies (among others) the d-separation condition $(X \perp\!\!\!\perp \{B, Z_3\}|\{Z_0, Z_2\})_G$ because all paths between X and $\{B, Z_3\}$ are blocked by $\{Z_0, Z_2\}$. However, G does not imply $(X \perp\!\!\!\perp \{B, Z_3\}|\{Z_0, Z_2, Y\})_G$ because the the path (X, Y, Z_3) contains a node (Y) drawing converging arrows which is also in the conditioning set $\{Z_0, Z_2, Y\}$.

An alternative test for d-separation has been devised by Lauritzen et al. (1990), based on the notion of ancestral graphs. To test for $(X \perp\!\!\!\perp Y|Z)_G$, delete from G all nodes except those in $\{X,\ Y,\ Z\}$ and their ancestors, connect by an edge every pair of nodes that share a common child, and remove all arrows from the arcs. $(X \perp\!\!\!\perp Y|Z)_G$ holds iff Z is a cutset of the resulting undirected graph, separating nodes of X from those of Y. Additional properties of DAGs and their applications to evidential reasoning in expert systems are discussed in Pearl (1988), Lauritzen & Spiegelhalter (1988), Spiegelhalter et al. (1993), and Pearl (1993a).

2.2 Graphs as models of interventions

The interpretation of DAGs as carriers of independence assumptions does not specifically mention causation, and DAGs displaying such assumptions can in fact be constructed for any ordering (not necessarily causal or chronological) of the variables. However, the main use of DAGs lies in their ability to portray causal, rather than statistical, associations. Causal models, assuming they are properly validated, are more informative than probability models because they also encode effects of actions. In other words, a joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions, such as those encountered in policy analysis and treatment management.

The connection between the causal and associational readings of DAGs is formed through the mechanism-based account of causation, which owes its roots to early works in econometrics (Frisch, 1938; Haavelmo, 1943; Simon, 1953). In this account, assertions about causal influences, such as those specified by the links in Figure 1, stand for autonomous physical mechanisms among the corresponding quantities, and these mechanisms are represented as functional relationships perturbed by random disturbances. In other words, each child-parent family in a DAG G represents a deterministic function

$$X_i = f_i(\mathbf{pa}_i, \epsilon_i), \qquad i = 1, ..., n \tag{3}$$

where $\mathbf{pa}_i$ are the parents of variable X_i in G, and ϵ_i, $1 \leq i \leq n$, are mutually independent, arbitrarily distributed random disturbances (Pearl & Verma, 1991). These disturbance terms represent independent exogenous factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables (thus violating the independence assumption), then that factor must enter the analysis as an unmeasured (or latent) variable, to be represented in the graph by a hollow node, such as Z_0 and B in Figure 1. For example, the causal assumptions conveyed by the model in Figure 1 correspond to the following set of equations:

$$\begin{array}{llll} Z_0 & = f_0(\epsilon_0) & Z_2 & = f_2(X, Z_1, \epsilon_2) \\ B & = f_B(Z_0, \epsilon_B) & Z_3 & = f_3(B, Z_2, \epsilon_3) \\ Z_1 & = f_1(Z_0, \epsilon_1) & Y & = f_Y(X, Z_2, Z_3, \epsilon_Y) \\ X & = f_X(Z_0, \epsilon_X) & & \end{array} \tag{4}$$

The equational model in (3) is the nonparametric analogue of the so-called structural equations model (Wright, 1921; Goldberger, 1973), with one exception: the functional form of the equations as well as the distribution of the disturbance terms will remain unspecified. The equality signs in structural equations convey the asymmetrical counterfactual relation of "is determined by", thus forming a clear correspondence between causal diagrams and Rubin's model of potential outcome (Rubin, 1974; Holland, 1988; Pratt & Schlaifer, 1988; Rubin, 1990). For example, the equation for Y states that regardless of what we currently observe about Y, and regardless of any changes that might occur in other equations, if $(X, Z_2, Z_3, \epsilon_Y)$ were to assume the values $(x, z_2, z_3, \epsilon_Y)$, respectively, Y would take on the value dictated by the function f_Y. Thus, the corresponding potential response variable in Rubin's model $Y_{(x)}$ (read: the value that Y would take if X were x) becomes a deterministic function of Z_2, Z_3 and ϵ_Y and can be considered a random variable whose distribution is determined by those of Z_2, Z_3 and ϵ_Y. The relation between graphical and counterfactual models is further analyzed in Appendix II and Pearl (1994a).

Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid \mathbf{pa}_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition that characterizes DAG models (see Eq. (2)). This occurs because each ϵ_i is independent on all nondescendants of X_i. However, the functional characterization $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ also provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select subset of functions, while keeping the others intact. Once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of the intervention can be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i. Such an intervention, which we call atomic, amounts to lifting X_i from the influence of the old functional mechanism $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $set(X_i = x_i)$, or $set(x_i)$ for short, amounts to removing the equation $X_i = f_i(\mathbf{pa}_i, \epsilon_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model thus created represents the system's behavior under the intervention $set(X_i = x_i)$ and, when solved for the distribution of X_j, yields the causal effect of X_i on X_j, denoted $P(x_j|\hat{x}_i)$. More generally, when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the model given in Eq. (3), one for each member of X, thus defining a new distribution over the remaining variables, which completely characterizes the effect of the intervention.[2] We therefore define:

DEFINITION 2. (causal effect) Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted $P(y|\hat{x})$, is a function from X to the space of probability distributions on Y. For each realization x of $X, P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the model (8) all equations corresponding to variables in X and substituting $X = x$ in the remaining equations. □

Clearly the graph corresponding to the reduced set of equations is an edge subgraph of G from which all arrows entering X have been pruned. We will denote this subgraph by $G_{\overline{X}}$.

An alternative (but operationally equivalent) account of intervention treats the force responsible for the intervention as a variable within the system (Pearl, 1993c). This is facilitated by representing the identity of the function f_i itself as a variable F_i and writing

$$X_i = I(\mathbf{pa}_i, F_i, \epsilon_i) \tag{5}$$

where I is a 3-argument function defined by

$$I(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

Thus, the impact of any external intervention that alters f_i can be represented graphically as an added parent node F_i of X_i, and the effect of such an intervention can be analyzed by Bayesian conditionalization, that is, by conditioning our probability on the added variable having attained the value f_i.

The effect of an atomic intervention $set(X_i = x_i')$ is encoded by adding to G a link $F_i \longrightarrow X_i$ (see Figure 2), where F_i is a new variable taking values in $\{set(x_i'),\ idle\}$, x_i' ranges over the domain of X_i, and $idle$ represents no intervention. Thus, the new parent set of X_i in the augmented network is $\mathbf{pa}_i' = \mathbf{pa}_i \cup \{F_i\}$, and it is related to

[2] An explicit translation of interventions to "wiping out" equations from the model was first proposed by Strotz & Wold (1960)and later used in Fisher (1970) and Sobel (1990). Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b). A related mathematical model, using event trees has been introduced by Robins (1986, pp. 1422–1425).

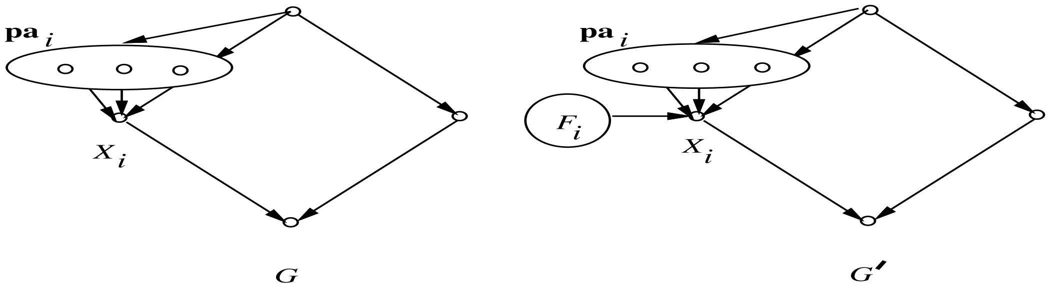

Figure 2: Representing external intervention F_i by an augmented network $G' = G \cup \{F_i \rightarrow X_i\}$.

X_i by the conditional probability

$$P(x_i \mid \mathbf{pa}'_i) = \begin{cases} P(x_i \mid \mathbf{pa}_i) & \text{if } F_i = idle \\ 0 & \text{if } F_i = set(x'_i) \text{ and } x_i \neq x'_i \\ 1 & \text{if } F_i = set(x'_i) \text{ and } x_i = x'_i \end{cases} \tag{6}$$

The effect of the intervention $set(x'_i)$ is to transform the original probability function $P(x_1, ..., x_n)$ into a new probability function $P(x_1, ..., x_n | \hat{x}'_i)$, given by

$$P(x_1, ..., x_n | \hat{x}'_i) = P'(x_1, ..., x_n \mid F_i = set(x'_i)) \tag{7}$$

where P' is the distribution specified by the augmented network $G' = G \cup \{F_i \rightarrow X_i\}$ and Eq. (6), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in G, we can construct an augmented probability function $P'(x_1, ..., x_n; F_1, ..., F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i's (taking values in their respective $set(x'_i)$), while the pre-intervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value *idle*.

Regardless of whether we represent interventions as a modification of an existing model (Definition 2) or as a conditionalization in an augmented model (Eq. (7), the result is a well-defined transformation between the pre-intervention and the post-intervention distributions. In the case of an atomic intervention $set(X_i = x'_i)$, this transformation can be expressed in a simple algebraic formula that follows immediately from Eq. (3) and Definition[3]:

$$P(x_1, ..., x_n | \hat{x}'_i) = \begin{cases} \frac{P(x_1, ..., x_n)}{P(x_i \mid \mathbf{pa}_i)} = \prod_{j \neq i} P(x_i | \mathbf{pa}_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{8}$$

This formula reflects the removal of the term $P(x_i \mid \mathbf{pa}_i)$ from the product of Eq. (2), since $\mathbf{pa}_i$ no longer influence X_i. Graphically, the removal of this term is equivalent to

[3]Eq. (8) can also be obtained from the G-computation formula of Robins (1986, p. 1423) and the Manipulation Theorem of Spirtes et al. (1993) (according to this source, such formula was "independently conjectured by Fienberg in a seminar in 1991"). Additional properties of the transformation defined in Eq. (8) are given in Pearl (1993b).

removing the links between $\mathbf{pa}_i$ and X_i while keeping the rest of the network intact. Clearly, then, an intervention $set(x_i)$ can affect only the descendants of X_i in G.

The immediate implication of Eq. (8) is that, given a causal diagram in which all parents of intervened variables are observable, one can infer post-intervention distributions from pre-intervention distributions; hence, under such assumptions we can estimate the effects of interventions from passive (i.e., nonexperimental) observations. The aim of this paper, however, is to derive causal effects in situations such as Figure 1, where some members of $\mathbf{pa}_i$ may be unobservable, thus preventing estimation of $P(x_i|\mathbf{pa}_i)$. The next two sections provide simple graphical tests for deciding when $P(x_j|\hat{x}_i)$ is estimable in a given model.

3 CONTROLLING CONFOUNDING BIAS

3.1 The back-door criterion

Assume we are given a causal diagram G together with nonexperimental data on a subset V_0 of observed variables in G and we wish to estimate what effect the intervention $set(X_i = x_i)$ would have on some response variable X_j. In other words, we seek to estimate $P(x_j|\hat{x}_i)$ from a sample estimate of $P(V_0)$.

The variables in $V_0 \backslash \{X_i, X_j\}$, are commonly known as concomitants (Cox, 1958, p. 48). In observational studies, concomitants are used to reduce confounding bias due to spurious correlations between treatment and response. The condition that renders a set Z of concomitants sufficient for identifying causal effect, also known as ignorability, has been given a variety of formulations, all requiring conditional independence judgments involving counterfactual variables (Rosenbaum & Rubin, 1983; Pratt & Schlaifer, 1988). In Pearl (1993b) it is shown that such judgments are equivalent to a simple graphical test, named the "back-door criterion", which can be applied directly to the causal diagram.[4]

DEFINITION 3. (back-door) A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if

(i) no node in Z is a descendant of X_i, and

(ii) Z blocks every path between X_i and X_j which contains an arrow into X_i.

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$. □

The name back-door echoes condition (ii), which requires that only paths with arrows pointing at X_i be d-blocked; these paths can be viewed as entering X_i through the back door. In Figure 3, for example, the sets $Z_1 = \{X_3, X_4\}$ and $Z_2 = \{X_4, X_5\}$ meet the back-door criterion, but $Z_3 = \{X_4\}$ does not because X_4 does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$.

[4]An equivalent, though more complicated, graphical criterion is given in Theorem 7.1 of Spirtes et al. (1993). An alternative criterion, using a single d-separation test will be established in Section 4 (see Eq. (25)).

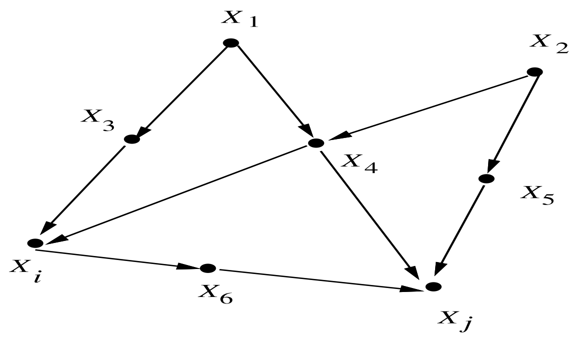

Figure 3: A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j|\hat{x}_i)$.

We summarize this finding in a theorem, after formally defining "identifiability".

DEFINITION 4. (identifiability) The causal effect of X on Y is said to be identifiable if the quantity $P(y|\hat{x})$ can be computed uniquely from any positive distribution of the observed variables. Identifiability means that $P(y|\hat{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution. □

Restricting identifiability to positive distributions substantially simplifies the analysis, as it avoids pathological cases associated with deterministic relationships (e.g., zero denominator in Eq. (8)). Extensions to some nonpositive distributions are feasible, but will not be treated here. Note that, to prove nonidentifiability, it is sufficient to present two sets of structural equations that induce identical distributions over observed variables but different causal effects.

THEOREM 1. If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z) \tag{9}$$

□

Eq. (9) represents the directly standardized adjustment for concomitants Z when X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983). Reducing ignorability conditions to the graphical criterion of Definition 3 replaces judgments about counterfactual dependencies with systematic procedures that can be applied to causal diagrams of any size and shape. The graphical criterion also enables the analyst to search for an optimal set of concomitants, namely, a set Z that minimizes measurement cost or sampling variability.

3.2 The front-door criteria

Condition (i) of Definition 3 reflects the prevailing practice that "the concomitant observations should be quite unaffected by the treatment" (Cox, 1958, p. 48). This subsection demonstrates how concomitants that *are* affected by the treatment can be used to facilitate causal inference. The emerging criterion, which we will name the front-door criterion, will constitute the second building block of the general test for identifying causal effects which will be formulated in Section 4.

Consider the diagram in Figure 4, which obtains from Figure 3 in case variables $X_1, \ldots, X_5$ are unobserved. Although Z does not satisfy any of the back-door conditions, measurements of Z can nevertheless enable consistent estimation of $P(y|\hat{x})$. This will be shown by reducing the expression for $P(y|\hat{x})$ to formulae computable from the observed distribution function $P(x, y, z)$.

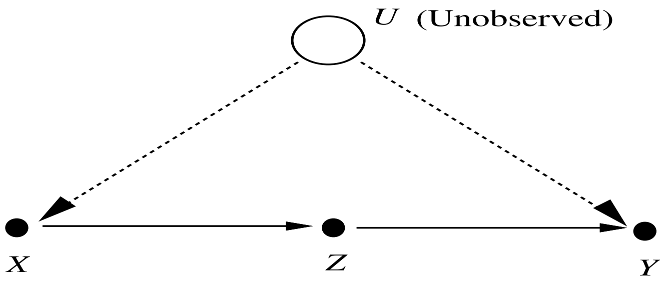

Figure 4: A diagram representing the front-door criterion.

The joint distribution associated with Figure 4 can be decomposed (Eq. (2)) into

$$P(x, y, z, u) = P(u)P(x|u)P(z|x)P(y|z, u) \tag{10}$$

From Eq. (8), the intervention $set(x)$ removes the factor $P(x|u)$ and induces the post-intervention distribution

$$P(y, z, u|\hat{x}) = P(y|z, u)P(z|x)P(u) \tag{11}$$

Summing over z and u gives

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_u P(y|z, u)P(u) \tag{12}$$

To eliminate u from the r.h.s. of Eq. (12), we use the two conditional independence assumptions encoded in the graph of Figure 4

$$\begin{aligned} P(u|z, x) &= P(u|x) & (13) \\ P(y|x, z, u) &= P(y|z, u) & (14) \end{aligned}$$

which yields the equality

$$\begin{aligned} \sum_u P(y|z, u)P(u) &= \sum_x \sum_u P(y|z, u)P(u|x)P(x) \\ &= \sum_x \sum_u P(y|x, z, u)P(u|x, z)P(x) \\ &= \sum_x P(y|x, z)P(x) & (15) \end{aligned}$$

and allows the reduction of Eq. (12) to the desired form:

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{16}$$

Since all factors on the r.h.s. of Eq. (16) are consistently estimable from nonexperimental data, it follows that $P(y|\hat{x})$ is estimable as well. Thus, we are in possession of an identifiable nonparametric estimand for the causal effect of an X on a Y whenever we can find a mediating variable Z that meets the conditions of Eqs. (13) and (14).

Eq. (16) can be interpreted as a two-step application of the back-door formula. In the first step we find the causal effect of X on Z and, since there is no back-door path from X to Z, we simply have

$$P(z|\hat{x}) = P(z|x)$$

Next, we compute the causal effect of Z on Y, which we can no longer equate with the conditional probability $P(y|z)$ because there is a back-door path $Z \leftarrow X \leftarrow U \rightarrow Y$ from Z to Y. However, since X blocks (d-separates) this path, X can play the role of a concomitant in the back-door criterion, which allows us to compute the causal effect of Z on Y in accordance with Eq. (9). Finally, we combine the two causal effects via

$$P(y|\hat{x}) = \sum_z P(y|\hat{z})P(z|\hat{x})$$

which reduces to Eq. (16).

We summarize this result by a theorem, after formally defining the assumptions.

DEFINITION 5. A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if

(i) Z intercepts all directed paths from X to Y.

(ii) There is no back-door path from X to Z.

(iii) All back-door paths from Z to Y are blocked by X. □

THEOREM 2. If Z satisfies the front-door criterion relative to (X, Y), and $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{17}$$

□

The conditions stated in Definition 5 are overly restrictive; some of the back-door paths excluded by conditions (ii) and (iii) can in fact be allowed, as long as they are blocked by some concomitants. For example, the variable Z_2 in Figure 1 satisfies a front-door-like criterion relative to (X, Z_3) by virtue of Z_1 blocking all back-door paths from X to Z_2 as well as those from Z_2 to Z_3. To allow the analysis of such intricate structures, including nested combinations of back-door and front-door conditions, a more powerful symbolic machinery will be introduced in Section 5, one that will sidestep algebraic manipulations such as those used in the derivation of Eq. (15). But first let us look at an example illustrating possible applications of the front-door condition.

3.3 Example: Smoking and the Genotype Theory

Consider the century-old debate on the relation between smoking (X) and lung cancer (Y) (Spirtes et al., 291–203, 1993). According to many, the tobacco industry has managed to stay anti-smoking legislation by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype (U) which involves inborn craving for nicotine.

The amount of tar (Z) deposited in a person's lungs is a variable that promises to meet the conditions listed in Definition 5 above, thus fitting the structure of Figure 4. To meet condition (i), we must assume that smoking cigarettes has no effect on the production of lung cancer except the one mediated through tar deposits. To meet conditions (ii) and (iii), we must assume that, even if a genotype is aggravating the production of lung cancer, it nevertheless has no effect on the amount of tar in the lungs except indirectly, through cigarette smoking. Finally, condition $P(x,z) > 0$ of Theorem 2 requires that we allow that high levels of tar in the lungs could be the result not only of cigarette smoking but also of other means (e.g., exposure to environmental pollutants) and that tar may be absent in some smokers (perhaps due to an extremely efficient tar-rejecting mechanism). Satisfaction of this last condition can be tested in the data.

To demonstrate how we can assess the degree to which cigarette smoking increases (or decreases) lung cancer risk, we will assume a hypothetical study in which the three variables, X, Y, and Z, were measured simultaneously on a large, randomly selected sample from the population. To simplify the exposition, we will further assume that all three variables are binary, taking on true (1) or false (0) values. A hypothetical data set from a study on the relations among tar, cancer, and cigarette smoking is presented in Table 1.

	Group Type	$P(x,z)$ Group Size (% of Population)	$P(Y=1\|x,z)$ % of Cancer Cases in Group
$X=0,\ Z=0$	Non-smokers, No tar	47.5	10
$X=1,\ Z=0$	Smokers, No tar	2.5	90
$X=0,\ Z=1$	Non-smokers, Tar	2.5	5
$X=1,\ Z=1$	Smokers, Tar	47.5	85

Table 1

It shows that 95% of smokers and 5% of non-smokers have developed high levels of tar in their lungs. Moreover, 81.51% of subjects with tar deposits have developed lung cancer, compared to only 14% among those with no tar deposits. Finally, within each of these two groups, tar and no tar, smokers show a much higher percentage of cancer than non-smokers.

These results seem to prove that smoking is a major contributor to lung cancer. However, the tobacco industry might argue that the table tells a different story – that smoking actually decreases, not increases, one's risk of lung cancer. Their argument goes as follows. If you decide to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you decide not to smoke. To evaluate the effect of tar deposits, we look separately at two groups, smokers and non-smokers. The table shows that tar deposits have a protective effect in both groups: in smokers, tar deposits lower cancer rates from 90% to 85%; in non-smokers, they lower cancer rates from 10% to 5%. Thus, regardless of whether I have a natural craving for nicotine, I should be seeking the protective effect of tar deposits in my lungs, and smoking offers a very effective means of acquiring them.

To settle the dispute between the two interpretations, we now apply the front-door formula (Eq. (17)) to the data in Table 1. We wish to calculate the probability that a randomly selected person will develop cancer under each of the following two actions: smoking (setting $X = 1$) or not smoking (setting $X = 0$).

Substituting the appropriate values of $P(y|x)$, $P(y|x, z)$, and $P(x)$ gives

$$\begin{aligned} P(Y = 1|set(X = 1)) &= .05(.10 \times .50 + .90 \times .50) + .95(.05 \times .50 + .85 \times .50) \\ &= .05 \times .50 + .95 \times .45 = .4525 \\ P(Y = 1|set(X = 0)) &= .95(.10 \times .50 + .90 \times .50) + .05(.05 \times .50 + .85 \times .50) \\ &= .95 \times .50 + .05 \times .45 = .4975 \end{aligned} \tag{18}$$

Thus, contrary to expectation, the data prove smoking to be somewhat beneficial to one's health.

The data in Table 1 are obviously unrealistic and were deliberately crafted so as to support the genotype theory. However, the purpose of this exercise was to demonstrate how reasonable qualitative assumptions about the workings of mechanisms, coupled with nonexperimental data, can produce precise quantitative assessments of causal effects. In reality, we would expect observational studies involving mediating variables to refute the genotype theory by showing, for example, that the mediating consequences of smoking, such as tar deposits, tend to increase, not decrease, the risk of cancer in smokers and non-smokers alike. The estimand of Eq. (17) could then be used for quantifying the causal effect of smoking on cancer.

4 A CALCULUS OF INTERVENTION

This section establishes a set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving (or verifying) claims about interventions. We will assume that we are given the structure of a causal diagram G in which some of the nodes are observable while the others remain unobserved. Our main problem will be to facilitate the syntactic derivation of causal effect expressions of the form $P(y|\hat{x})$, where X and Y stand for any subsets of observed variables. By derivation we mean step-wise reduction of the expression $P(y|\hat{x})$ to an equivalent expression involving standard probabilities of observed quantities. Whenever such reduction is feasible, the causal effect of X on Y is identifiable (see Definition 4).

4.1 Preliminary notation

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$ (see Figure 5 for illustration). Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

4.2 Inference rules

The following theorem states the three basic inference rules of the proposed calculus. Proofs are provided in the appendix.

THEOREM 3. Let G be the directed acyclic graph associated with a causal model as defined in Eq. (3), and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W we have:

Rule 1 Insertion/deletion of observations

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{19}$$

Rule 2 Action/observation exchange

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{20}$$

Rule 3 Insertion/deletion of actions

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{21}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of the inference rules above follows from the basic interpretation of the "$\hat{x}$" operator as a replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in Spirtes et al. (1993)) which supports all three rules.

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $set(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms (see Eq. (3)).

Rule 2 provides a condition for an external intervention $set(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $set(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $set(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$).

COROLLARY 1. A causal effect q: $P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 3, which reduces q into a standard (i.e., hat-free) probability expression involving observed quantities. $\square$

Whether the three rules above are sufficient for deriving all identifiable causal effects remains an open question. However, the task of finding a sequence of transformations (if such exists) for reducing an arbitrary causal effect expression can be systematized and executed by efficient algorithms [Galles 1994, unpublished report]. As the next subsection illustrates, symbolic derivations using the hat notation are much more convenient than algebraic derivations that aim at eliminating latent variables from standard probability expressions (as in Section 3.2).

4.3 Symbolic derivation of causal effects: an example

We will now demonstrate how Rules 1-3 can be used to derive all causal effect estimands in the structure of Figure 4 above. Figure 5 displays the subgraphs that will be needed for the derivations that follow.

Task-1, compute $P(z|\hat{x})$

This task can be accomplished in one step, since G satisfies the applicability condition for Rule 2; namely, $X \perp\!\!\!\perp Z$ in $G_{\underline{X}}$ (because the path $X \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by the converging arrows at Y) and we can write

$$P(z|\hat{x}) = P(z|x) \tag{22}$$

Task-2, compute $P(y|\hat{z})$

Here we cannot apply Rule 2 to exchange $\hat{z}$ with z because $G_{\underline{Z}}$ contains a back-door path from Z to $Y : Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by measuring variables (such as X) that reside on that path. This involves conditioning and summing over all values of X,

$$P(y|\hat{z}) = \sum_x P(y|x, \hat{z})P(x|\hat{z}) \tag{23}$$

We now have to deal with two expressions involving $\hat{z}$, $P(y|x, \hat{z})$ and $P(x|\hat{z})$. The latter can be readily computed by applying Rule 3 for action deletion:

$$P(x|\hat{z}) = P(x) \;\text{ if }\; (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}} \tag{24}$$

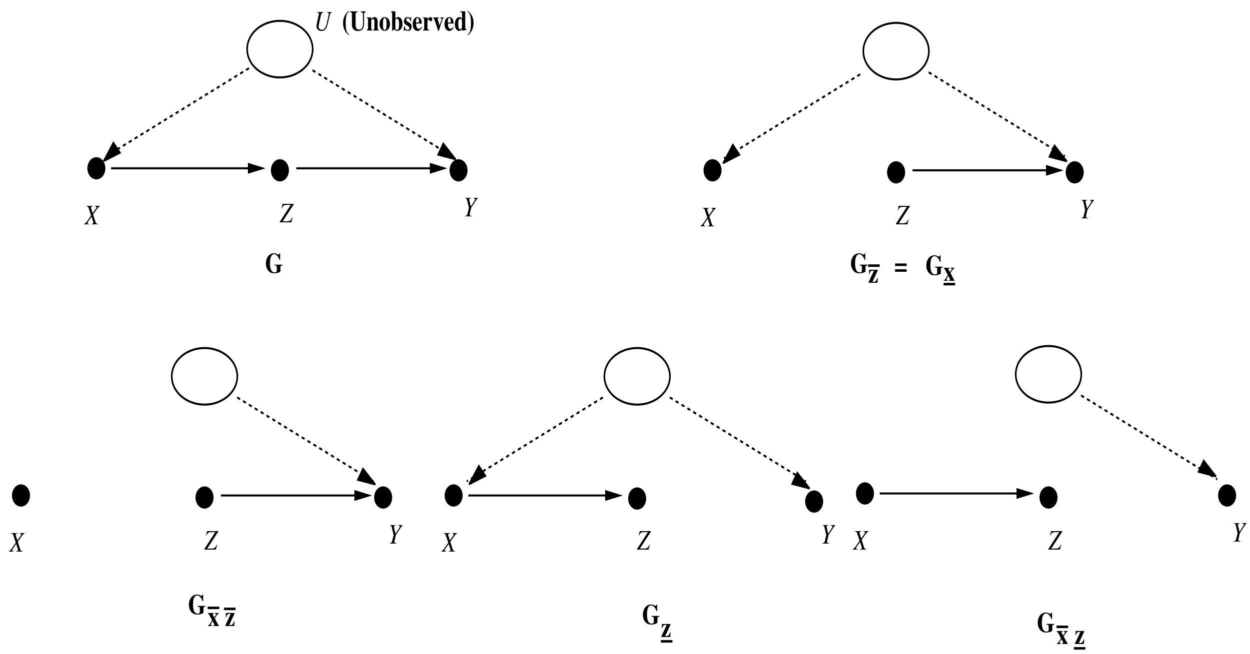

Figure 5: Subgraphs of G used in the derivation of causal effects.

since X and Z are d-separated in $G_{\overline{Z}}$. (Intuitively, manipulating Z should have no effect on X, because Z is a descendant of X in G.) To reduce the former, $P(y|x,\hat{z})$, we consult Rule 2:

$$P(y|x,\hat{z}) = P(y|x,z) \text{ if } (Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}} \tag{25}$$

noting that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write Eq. (23) as

$$P(y|\hat{z}) = \sum_x P(y|x,z)P(x) = E_x P(y|x,z) \tag{26}$$

which is a special case of the back-door formula (Eq. (9)). The legitimizing condition, $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$, offers yet another graphical test for the ignorability condition of Rosenbaum & Rubin (1983).

Task-3, compute $P(y|\hat{x})$

Writing

$$P(y|\hat{x}) = \sum_z P(y|z,\hat{x})P(z|\hat{x}) \tag{27}$$

we see that the term $P(z|\hat{x})$ was reduced in Eq. (22) but that no rule can be applied to eliminate the "hat" symbol ˆ from the term $P(y|z,\hat{x})$. However, we can add a "hat" symbol to this term via Rule 2

$$P(y|z,\hat{x}) = P(y|\hat{z},\hat{x}) \tag{28}$$

since the applicability condition $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\underline{Z}}}$, holds true (see Figure 5). We can now delete the action $\hat{x}$ from $P(y|\hat{z}, \hat{x})$ using Rule 3, since $Y \perp\!\!\!\perp X|Z$ holds in $G_{\overline{X}\overline{Z}}$. Thus, we have

$$P(y|z, \hat{x}) = P(y|\hat{z}) \tag{29}$$

which was calculated in Eq. (26). Substituting Eqs. (26), (29), and (22) back into Eq. (27) finally yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{30}$$

which is identical to the front-door formula of Eq. (16).

Task-4, compute $P(y, z|\hat{x})$

$$P(y, z|\hat{x}) = P(y|z, \hat{x})P(z|\hat{x})$$

The two terms on the r.h.s. were derived before in Eqs. (22) and (29), from which we obtain

$$\begin{aligned} P(y, z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x) \textstyle\sum_{x'} P(y|x', z)P(x') \end{aligned} \tag{31}$$

Task-5, compute $P(x, y|\hat{z})$

$$\begin{aligned} P(x, y|\hat{z}) &= P(y|x, \hat{z})P(x|\hat{z}) \\ &= P(y|x, z)P(x) \end{aligned} \tag{32}$$

The first term on the r.h.s. is obtained by Rule 2 (licensed by $G_{\underline{Z}}$) and the second term by Rule 3 (as in Eq. (24)).

that in all the derivations the graph G has provided both the license for applying the inference rules and the guidance for choosing the right rule to apply.

4.4 Causal inference by surrogate experiments

Suppose we wish to learn the causal effect of X on Y when $P(y|\hat{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomized experiment. The question arises whether $P(y|\hat{x})$ can be identified by randomizing a surrogate variable Z, which is easier to control than X. For example, if we are interested in assessing the effect of cholesterol levels (X) on heart disease (Y), a reasonable experiment to conduct would be to control subjects' diet (Z), rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $P(y|\hat{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 3 it can be shown that the following conditions are sufficient for admitting a surrogate variable Z: (i) X intercepts all directed paths from Z to Y, and, (ii) $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$. Indeed, if condition (i) holds, we can write $P(y|\hat{x}) = P(y|\hat{x},\ \hat{z})$, because $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\overline{Z}}}$. But $P(y|\hat{x},\ \hat{z})$ stands for the causal effect of X on Y in a model governed by $G_{\overline{Z}}$

which, by condition (ii), is identifiable. Translated to our cholesterol example, these condition require that there be no direct effect of diet on heart conditions and no confounding effect between cholesterol levels and heart disease, unless we can measure an intermediate variable between the two.

Figures 8(e) and 8(h) below illustrate models in which both conditions hold. For Figure 8(e), for example, we obtain this estimand

$$P(y|\hat{x}) = P(y|x,\ \hat{z}) = P(y,\ x|\hat{z})/P(x|\hat{z}) \tag{33}$$

This can be established directly by first applying Rule 3 to add $\hat{z}$,

$$P(y|\hat{x}) = P(y|\hat{x},\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$$

then applying Rule 2 to exchange $\hat{x}$ with x:

$$P(y|\hat{x},\ \hat{z}) = P(y|x,\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}\overline{Z}}}$$

According to Eq. (33), only one level of Z suffices for the identification of $P(y|\hat{x})$, for any values of y and x. In other words, Z need not be varied at all, just held constant by external means, and, if the assumptions embodied in G are valid, the r.h.s. of Eq. (33) should attain the same value regardless of the level at which Z is being held constant. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X. For example, if we are interested in the difference $E(Y|\hat{x}) - E(Y|\hat{x}')$, where x and x' are two treatment levels, then we should choose two values z and z' of Z which maximize the number of samples in x and x', respectively, and estimate

$$E(Y|\hat{x}) - E(Y|\hat{x}') = E(Y|x, \hat{z}) - E(Y|x', \hat{z}')$$

5 GRAPHICAL TESTS OF IDENTIFIABILITY

Figure 6 shows simple diagrams in which $P(y|\hat{x})$ cannot be identified due to the presence of a bow pattern, i.e., a confounding arc (dashed) embracing a causal link between X and Y. A confounding arc represents the existence in the diagram of a back-door path that contains only unobserved variables and has no converging arrows. For example, the path X, Z_0, B, Z_3 in Figure 1 can be represented as a confounding arc between X and Z_3. A bow-pattern represents an equation $Y = f_Y(X, U, \epsilon_Y)$ where U is unobserved and dependent on X. Such an equation does not permit the identification of causal effects since any portion of the observed dependence between X and Y may always be attributed to spurious dependencies mediated by U.

The presence of a bow-pattern prevents the identification of $P(y|\hat{x})$ even when it is found in the context of a larger graph, as in Figure 6(b). This is in contrast to linear models, where the addition of an arc to a bow-pattern can render $P(y|\hat{x})$ identifiable. For example, if Y is related to X via a linear relation $Y = bX + U$, where U is an unobserved disturbance possibly correlated with X, then $b = \frac{\partial}{\partial x}E(Y|\hat{x})$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure (that is, finding a variable Z that is correlated with X but not with U) would facilitate the computation of $E(Y|\hat{x})$

via the instrumental-variable formula (Bowden & Turkington, 1984; Angrist et al., 1993):

$$b \triangleq \frac{\partial}{\partial x} E(Y|\hat{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}} \tag{34}$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern (Figure 6(b)) does not permit the identification of $P(y|\hat{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment (Z) is randomized (hence, no link enters Z), but compliance is imperfect. The confounding arc between X and Y in Figure 6(b) represents unmeasurable factors which influence both subjects' choice of treatment (X) and subjects' response to treatment (Y). In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $P(y|\hat{x})$ without making additional assumptions on the nature of the interactions between compliance and response, as is done, for example, in the general approach to instrumental variables developed in Angrist et al. (1993) and Imbens & Angrist (1994). While the added arc $Z \rightarrow X$ permits us to calculate bounds on $P(y|\hat{x})$ (Robins, 1989, Sec. 1g; Manski, 1990) and the upper and lower bounds may even coincide for certain types of distributions $P(x, y, z)$ (Balke & Pearl, 1994), there is no way of computing $P(y|\hat{x})$ for every positive distribution $P(x, y, z)$, as required by Definition 4.

In general, the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects in nonparametric models. This is because such addition reduces the set of d-separation conditions carried by the diagram and, hence, if a causal effect derivation fails in the original diagram, it is bound to fail in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented diagram (by a sequence of symbolic transformations, as in Corollary 1) would succeed in the original diagram.

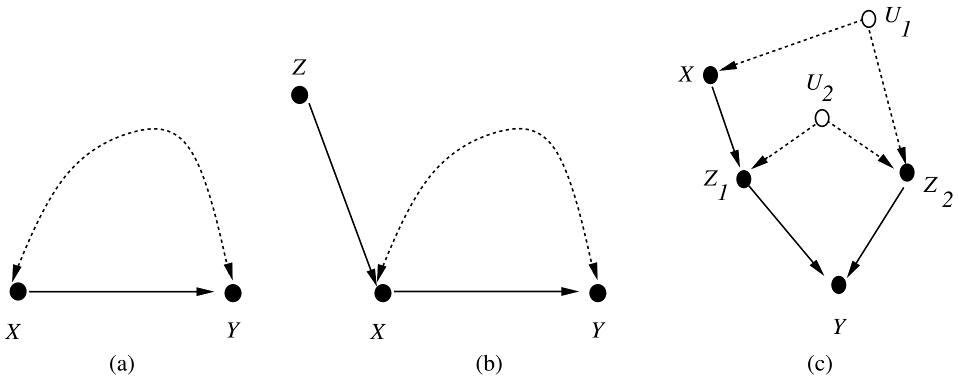

Figure 6: (a) A bow-pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y|\hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bow-less graph still prohibiting the identification of $P(y|\hat{x})$.

Our ability to compute $P(y|\hat{x})$ for pairs (x, y) of singleton variables does not ensure our ability to compute joint distributions, such as $P(y_1, y_2|\hat{x})$. Figure 6(c), for example, shows a causal diagram where both $P(z_1|\hat{x})$ and $P(z_2|\hat{x})$ are computable, but $P(z_1, z_2|\hat{x})$ is not. Consequently, we cannot compute $P(y|\hat{x})$. Interestingly, this diagram is the smallest graph that does not contain a bow-pattern and still presents an uncomputable causal effect.

Another interesting feature demonstrated by Figure 5(c) is that computing the effect of a joint intervention is often easier than computing the effects of its constituent singleton interventions.[5] Here, it is possible to compute $P(y|\hat{x}, \hat{z}_2)$ and $P(y|\hat{x}, \hat{z}_1)$, yet there is no way of computing $P(y|\hat{x})$. For example, the former can be evaluated by invoking Rule 2 in $G_{\overline{X}\underline{Z_2}}$, giving

$$P(y|\hat{x}, \hat{z}_2) = \sum_{z_1} P(y|z_1, \hat{x}, \hat{z}_2) P(z_1|\hat{x}, \hat{z}_2) = \sum_{z_1} P(y|z_1, x, z_2) P(z_1|x) \qquad (35)$$

However, Rule 2 cannot be used to convert $P(z_1|\hat{x}, z_2)$ into $P(z_1|x, z_2)$ because, when conditioned on Z_2, X and Z_1 are d-connected in $G_{\underline{X}}$ (through the dashed lines). A general approach to computing the effect of joint interventions is developed in Pearl & Robins (1995).

5.1 Identifying models

Figure 7 shows simple diagrams in which the causal effect of X on Y, is identifiable. Such models are called identifying because their structures communicate a sufficient number of assumptions (missing links) to permit the identification of the target quantity $P(y|\hat{x})$. Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs (dashed). Every causal diagram with latent variables can be converted to an equivalent diagram involving measured variables interconnected by arrows and confounding arcs. This conversion corresponds to substituting out all latent variables from the structural equations of Eq. (3) and then constructing a new diagram by connecting any two variables X_i and X_j by (1) an arrow from X_j to X_i whenever X_j appears in the equation for X_i and (2) a confounding arc whenever the same ϵ term appears in both f_i and f_j. The result is a diagram in which all unmeasured variables are exogenous and mutually independent.

Several features should be noted from examining the diagrams in Figure 7.

1. Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $P(y|\hat{x})$ will still be identified in any edge-subgraph of the diagrams shown in Figure 7. Likewise, the introduction of mediating observed variables onto any edge in a causal graph can assist, but never impede, the identifiability of any causal effect. Therefore, $P(y|\hat{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 7.

[5]This was brought to my attention by James Robins, who has worked out many of these computations in the context of sequential treatment management. Eq. (35) for example, can be obtained from Robin's G-computation algorithm (Robins, p. 1423, 1986).

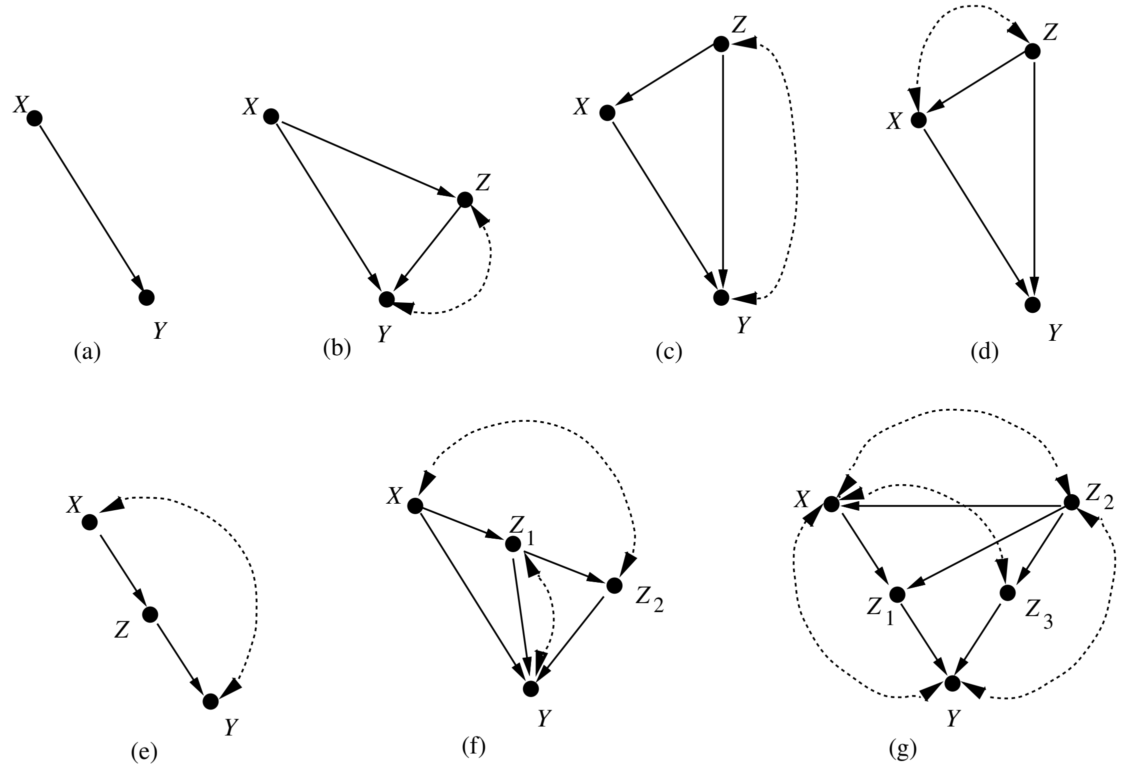

Figure 7: Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

2. The diagrams in Figure 7 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $P(y|\hat{x})$ no longer identifiable.

3. Although most of the diagrams in Figure 7 contain bow-patterns, none of these patterns emanates from X (as is the case in Figure 8(a) and (b) below). In general, a necessary condition for the identifiability of $P(y|\hat{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y.

4. Diagrams (a) and (b) in Figure 7 contain no back-door paths between X and Y, and thus represent experimental designs in which there is no confounding bias between the treatment (X) and the response (Y) (i.e., X is strongly ignorable relative to Y (Rosenbaum & Rubin, 1983); hence, $P(y|\hat{x}) = P(y|x)$. Likewise, diagrams (c) and (d) in Figure 7 represent designs in which observed covariates, Z, block every back-door path between X and Y (i.e., X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983); hence, $P(y|\hat{x})$ is obtained by standard adjustment for Z (as in Eq. (9)):

$$P(y|\hat{x}) = \sum_z P(y|x,z)P(z)$$

5. For each of the diagrams in Figure 7, we can readily obtain a formula for $P(y|\hat{x})$, by using symbolic derivations patterned after those in Section 4.3. The derivation is often guided by the graph topology. For example, diagram (f) in Figure 7 dictates the following derivation. Writing

$$P(y|\hat{x}) = \sum_{z_1,z_2} P(y|z_1, z_2, \hat{x}) P(z_1, z_2|\hat{x})$$

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of diagram (e), with (Z_1, Z_2) replacing (Z, Y), respectively. Thus, $P(z_1, z_2|\hat{x})$ can be obtained from (23) and (30). Likewise, the term $P(y|z_1, z_2, \hat{x})$ can be reduced to $P(y|z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\underline{X}}}$. Thus, we have

$$P(y|\hat{x}) = \sum_{z_1,z_2} P(y|z_1, z_2, x)\ P(z_1|x) \sum_{x'} P(z_2|z_1, x')\ P(x') \tag{36}$$

Applying a similar derivation to diagram (g) of Figure 7 yields

$$P(y|\hat{x}) = \sum_{z_1}\sum_{z_2}\sum_{x'} P(y|z_1,\ z_2,\ x')P(x')P(z_1|z_2,\ x)P(z_2) \tag{37}$$

Note that the variable Z_3 does not appear in the expression above, which means that Z_3 need not be measured if all one wants to learn is the causal effect of X on Y.

6. In diagrams (e), (f), and (g) of Figure 7, the identifiability of $P(y|\hat{x})$ is rendered feasible through observed covariates, Z, that are affected by the treatment X (i.e., Z being descendants of X). This stands contrary to the warning, repeated in most of the literature on statistical experimentation, to refrain from adjusting for concomitant observations that are affected by the treatment (Cox, 1958; Rosenbaum, 1984; Pratt & Schlaifer, 1988; Wainer, 1989). It is commonly believed that if a concomitant Z is affected by the treatment, then it must be excluded from the analysis of the total effect of the treatment (Pratt & Schlaifer, 1988). The reason given for the exclusion is that the calculation of total effects amounts to integrating out Z, which is functionally equivalent to omitting Z to begin with. Diagrams (e), (f), and (g) show cases where one wants to learn the total effects of X and, still, the measurement of concomitants that are affected by X (e.g., Z, or Z_1) is necessary. However, the adjustment needed for such concomitants is nonstandard, involving two or more stages of the standard adjustment of Eq. (9), (see Eqs. (16), (36), and (37)).

7. In diagrams (b), (c), and (f) of Figure 7, Y has a parent whose effect on Y is not identifiable yet the effect of X on Y is identifiable. This demonstrates that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y.

5.2 Nonidentifying models

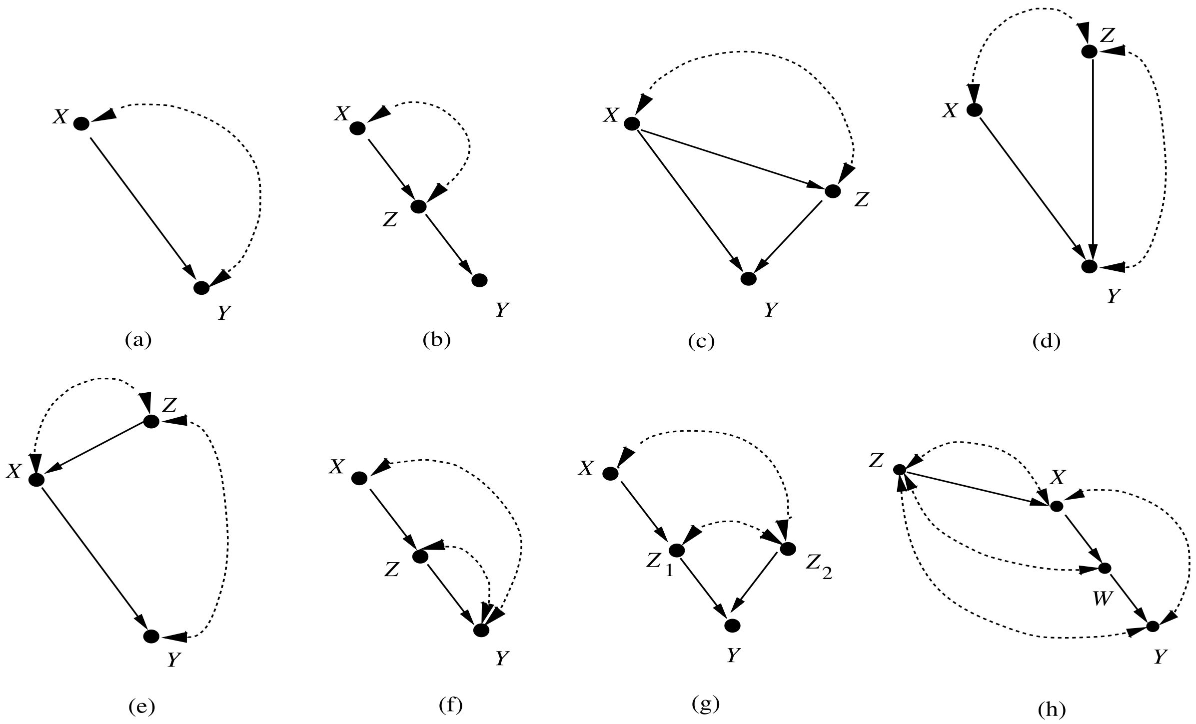

Figure 8: Typical models in which $P(y|\hat{x})$ is not identifiable.

Figure 8 presents typical diagrams in which the total effect of X on Y, $P(y|\hat{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

1. All graphs in Figure 8 contain unblockable back-door paths between X and Y, that is, paths ending with arrows pointing to X which cannot be blocked by observed nondescendants of X. The presence of such a path in a graph is, indeed, a necessary test for nonidentifiability (see Theorem 1). It is not a sufficient test, though, as is demonstrated by Figure 7(e), in which the back-door path (dashed) is unblockable and yet $P(y|\hat{x})$ is identifiable.

2. A sufficient condition for the nonidentifiability of $P(y|\hat{x})$ is the existence of a confounding path between X and any of its children on a path from X to Y, as shown in Figure 8(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 8 as an edge-subgraph.

3. Graph (g) in Figure 8 (same as 6(c)) demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $P(z_1|\hat{x})$, $P(z_2|\hat{x})$, $P(y,|\hat{z}_1)$, and $P(y|\hat{z}_2)$, but not $P(y|\hat{x})$. This is one of the main differences between nonparametric and linear models; in the latter, all causal effects can be determined from the structural coefficients, each coefficient representing the causal effect of one variable on its immediate successor.

6 DISCUSSION

The basic limitation of the methods proposed in this paper is that the results must rest on the causal assumptions shown in the graph, and that these cannot usually be tested in observational studies. In a related paper (Pearl, 1994a) we show that some of the assumptions, most notably those associated with instrumental variables (Figure 6 (b)) are subject to falsification tests.[6] Moreover, considering that any causal inferences from observational studies must ultimately rely on some kind of causal assumptions, the methods described in this paper offer an effective language for making those assumptions precise and explicit, so they can be isolated for deliberation or experimentation and, once validated, be integrated with statistical data.

A second limitation concerns an assumption inherent in identification analysis, namely, that the sample size is so large that sampling variability may be ignored. The mathematical derivation of causal-effect estimands should therefore be considered a first step toward supplementing these estimands with confidence intervals and significance levels, as in traditional analysis of controlled experiments. We should remark, though, that having obtained nonparametric estimands for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in Eq. (16) can be converted to the product $E(Y|\hat{x}) = R_{xz}\beta_{zy\cdot x}x$ where $\beta_{zy\cdot x}$ is the standardized regression coefficient, and the estimation problem reduces to that of estimating regression coefficients (e.g., by least-squares). More sophisticated estimation techniques, can be found in Rubin (1978), Robins (1989, Sec. 17), and Robins et al. (1992, pp. 331–333).

Several extensions of the methods proposed in this paper are noteworthy. First, the analysis of atomic interventions can be generalized to complex policies in which a set X of treatment variables is made to respond in a specified way to some set Z of covariates, say through a functional relationship $X = g(Z)$ or through a stochastic relationship whereby X is set to x with probability $P^*(x|z)$. In Pearl (1994b) it is shown that computing the effect of such policies is equivalent to computing the expression $P(y|\hat{x}, z)$.

A second extension concerns the use of the intervention calculus (Theorem 3) in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in term of "wiping out" equations from the model (Definition 2) still carries over to nonrecursive systems (Strotz & Wold, 1960; Sobel, 1990), but then two issues must be addressed. First, the analysis of identification must ensure the stability of the remaining submodels (Fisher, 1970). Second, the d-separation criterion for DAGs must be extended to cover cyclic graphs as well. The validity of d-separation has been established for nonrecursive linear models and extended, using an augmented graph, to any arbitrary set of stable equations

[6]The testable implications of the model of Figure 6 (b) can be expressed in a simple inequality (Pearl, 1994c): $\max_x \sum_y \max_z P(y, x|z) \leq 1$.

(Spirtes, 1994). However, the computation of causal effect estimands will be harder in cyclic networks, because symbolic reduction of $P(y|\hat{x})$ to hat-free expressions may require the solution of nonlinear equations.

Finally, a few comments regarding the notation introduced in this paper. Traditionally, statisticians have approved of only one method of combining subject-matter considerations with statistical data: the Bayesian method of assigning subjective priors to distributional parameters. To incorporate causal information within the Bayesian framework, plain causal statements such as "Y is affected by X" must be converted into sentences capable of receiving probability values, e.g., counterfactuals. Indeed, this is how Rubin's model has achieved statistical legitimacy: causal judgments are expressed as constraints on probability functions involving counterfactual variables (see Appendix II).

Causal diagrams offer an alternative language for combining data with causal information. This language simplifies the Bayesian route by accepting plain causal statements as its basic primitives. These statements, which merely identify whether a causal connection between two variables of interest exists, are commonly used in natural discourse and provide a natural way for scientists to communicate experience and organize knowledge. It can be anticipated, therefore, that the language of causal graphs will find applications in problems requiring substantial use of subject-matter considerations.

The language is not new. The use of diagrams and structural equations models to convey causal information has been quite popular in the social sciences and econometrics. Statisticians, however, have generally found these models suspect, perhaps because social scientists and econometricians have failed to provide an unambiguous definition of the empirical content of their models, that is, to specify the experimental conditions, however hypothetical, whose outcomes would be constrained by a given structural equation. As a result, even such basic notions as "structural coefficients" or "missing links" become the object of serious controversy (Freedman, 1987) and conflicting interpretations (Wermuth, 1992; Whittaker, 1990, p. 302; Cox & Wermuth, 1993).

To a large extent, this history of controversy and miscommunication stems from the absence of an adequate mathematical notation for defining basic notions of causal modeling. For example, standard probabilistic notation cannot express the empirical content of the coefficient b in the structural equation $Y = bX + \epsilon_Y$ even if one is prepared to assume that ϵ_Y (an unobserved quantity) is uncorrelated with X. Nor can any probabilistic meaning be attached to the analyst's excluding from the equation certain variables that are highly correlated with X or Y.

The notation developed in this paper gives these notions a clear empirical interpretation, because it permits one to specify precisely what is being held constant and what is merely measured in a controlled experiment. (The need for this distinction was recognized by many researchers, most notably Pratt & Schlaifer (1988) and Cox (1992)). The meaning of b is simply $\frac{\partial}{\partial x} E(Y|\hat{x})$, namely, the rate of change (in x) of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ϵ_Y and X are correlated (e.g., via another equation: $X = aY + \epsilon_X$) and, moreover, the notion of randomization need not be invoked. Likewise, the analyst's decision as to which variables should be included in a given equation can be based on a hypothetical controlled experiment: A variable Z is excluded from the equation for Y if it has no influence on Y when all other variables, S_{YZ}, are held constant, that is, $P(y|\hat{z}, \hat{s}_{YZ}) = P(y|\hat{s}_{YZ})$. Specifically, variables that are excluded from the equation $Y = bX + \epsilon_Y$ are not conditionally independent of Y given measurements of X, but rather causally irrelevant to Y given settings of X. The operational meaning of the so called "disturbance term", ϵ_Y, is likewise demystified: ϵ_Y is defined as the difference $Y - E(Y|\hat{s}_Y)$; two disturbance terms, ϵ_X and ϵ_Y, are correlated if $P(y|\hat{x}, \hat{s}_{XY}) \neq P(y|x, \hat{s}_{XY})$, and so on.

The distinctions provided by the "hat" notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers. Moreover, since most scientific knowledge is organized around the operation of "holding X fixed," rather than "conditioning on X," the notation and calculus developed in this paper should provide an effective means for scientists to communicate subject-matter information, and to infer its logical consequences.

APPENDIX I: Proof of Theorem 3

1. Rule 1 follows from the fact that deleting equations from the model in Eq. (8) results, again, in a recursive set of equations in which all ϵ terms are mutually independent. The d-separation condition is valid for any recursive model, hence it is valid for the submodel resulting from deleting the equations for X. Finally, since the graph characterizing this submodel is given by $G_{\overline{X}}$, $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$ implies the conditional independence $P(y|\hat{x}, z, w) = P(y|\hat{x}, w)$ in the post-intervention distribution.

2. The graph $G_{\overline{X}\underline{Z}}$ differs from $G_{\overline{X}}$ only in lacking the arrows emanating from Z, hence it retains all the back-door paths from Z to Y that can be found in $G_{\overline{X}}$. The condition $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$ ensures that all back-door paths from Z to Y (in $G_{\overline{X}}$) are blocked by $\{X, W\}$. Under such conditions, setting $(Z = z)$ or conditioning on $Z = z$ has the same effect on Y. This can best be seen from the augmented diagram $G'_{\overline{X}}$, to which the intervention arcs $F_Z \rightarrow Z$ were added, where F_z stands of the external intervention as in Figure 2. If all back-door paths from F_Z to Y are blocked, the remaining paths from F_Z to Y must go through the children of Z, hence these paths will be blocked by Z. The implication is that Y is independent of F_Z given Z, which means that the observation $Z = z$ cannot be distinguished from the intervention $F_Z = set(z)$.

3. (After D. Galles) Consider the augmented diagram $G'_{\overline{X}}$ to which the intervention arcs $F_z \rightarrow Z$ are added. If $(F_Z \perp\!\!\!\perp Y|W, X)_{G'_{\overline{X}}}$, then $P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w)$. If $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$, and $(F_Z \not\perp\!\!\!\perp Y|W, X)_{G'_{\overline{X}}}$, there must be an unblocked

path from a member $F_{Z'}$ of F_Z to Y that passes either through a head-to-tail junction at Z', or a head-to-head junction at Z'. If there is such a path, let P be the shortest such path. We will show that P will violate some premise, or there exists a shorter path, either of which leads to a contradiction.

If the junction is head-to-tail, that means that $(Y \not\perp\!\!\!\perp Z'|W, X)_{G'_{\overline{X}}}$, but $(Y \perp\!\!\!\perp Z'|W, X)_{G'_{\overline{X}\ \overline{Z(W)}}}$. So, there must be an unblocked path from Y to Z' that passes through some member Z'' of $Z(W)$ in either a head-to-head or a tail-to-head junction. This is impossible. If the junction is head-to-head, then some descendant of Z'' must be in W for the path to be unblocked, but then Z'' would not be in $Z(W)$. If the junction is tail-to-head, there are two options : either the path from Z' to Z'' ends in a arrow pointing to Z'', or an arrow pointing away from Z''. If it ends in an arrow pointing away from Z'', then there must be a head-to-head junction along the path from Z' to Z''. In that case, for the path to be unblocked, W must be a descendant of Z'', but then Z'' would not be in $Z(W)$. If it ends in an arrow pointing to Z'', then there must be an unblocked path from Z'' to Y in $G_{\overline{X}}$ that is blocked in $G_{\overline{X}\ \overline{Z(W)}}$. If this is true, then there is an unblocked path from $F_{Z''}$ to Y that is shorter than P, the shortest path.

If the junction through Z' is head-to-head, then either Z' is in $Z(W)$, in which case that junction would be blocked, or there is an unblocked path from Z' to Y in $G_{\overline{X},\ \overline{Z(W)}}$ that is blocked in $G_{\overline{X}}$. Above, we proved that this could not occur. So $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\ \overline{Z(W)}}}$ implies $(F_Z \perp\!\!\!\perp Y|W, X)_{G'_{\overline{X}}}$, and thus $P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w)$.

APPENDIX II: Graphs, structural equations, and counterfactuals

This paper uses two representations of causal models: graphs and structural equations. By now, both representations have been considered controversial for almost a century. On the one hand, economists and social scientists have embraced these modeling tools, but they continue to debate the empirical content of the symbols they estimate and manipulate; as a result, the use of structural models in policy-making contexts is often viewed with suspicion. Statisticians, on the other hand, reject both representations as problematic (if not meaningless) and instead resort to the Neyman-Rubin counterfactual notation whenever they are pressed to communicate causal information. This appendix presents an explication that unifies these three representation schemes in order to uncover commonalities, mediate differences, and make the causal-inference literature more generally accessible.

The primitive object of analysis in Rubin's counterfactual framework is the unit-based response variable, denoted $Y(x, u)$ or $Y_x(u)$, read: "the value that Y would obtain in unit u, had X been x". This variable has natural interpretation in structural equations models. Consider a set T of equations

$$X_i = f_i(PA_i, U_i) \quad i = 1, \dots, n \tag{38}$$

where the U_i stand for latent exogenous variables (or disturbances), and the PA_i are the explanatory (observed) variables in the ith equation (pa_i is a realization of PA_i). (38) is similar to (3), except we no longer insist on the equations being recursive or on the U_i's being independent. Let U stand for the vector $(U_1, \ldots, U_n)$, let X and Y be two disjoint subsets of observed variables, and let T_x be the submodel created by replacing the equations corresponding to variables in X with $X = x$, as in Definition 2. The structural interpretation of $Y(x, u)$ is given by

$$Y(x, u) \triangleq Y_{T_x}(u) \tag{39}$$

namely, $Y(x, u)$ is the (unique) solution of Y under the realization $U = u$ in the submodel T_x of T. While the term *unit* in the counterfactual literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on, which are represented as components of the vector u in structural modeling. Eq. (39) forms a connection between the opaque English phrase "the value that Y would obtain in unit u, had X been x" and the physical processes that transfer changes in X into changes in Y. The formation of the submodel T_x represents a minimal change in model T needed for making x and u compatible; such a change could result either from external intervention or from a natural yet unanticipated eventuality.

Given this interpretation of $Y(x, u)$, it is instructive to contrast the methodologies of causal inference in the counterfactual and the structural frameworks. If U is treated as a random variable, then the value of the counterfactual $Y(x, u)$ becomes a random variable as well, denoted as $Y(x)$ or Y_x. The counterfactual analysis proceeds by imagining the observed distribution $P^*(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects, written $P(y|\hat{x})$ in the structural analysis, are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y(x) = y)$. The new entities $Y(x)$ are treated as ordinary random variables that are connected to the observed variables via consistency constraints (Robins, 1987) such as

$$X = x \implies Y(x) = Y \tag{40}$$

and a set of conditional independence assumptions which the investigator must supply to endow the augmented probability, P^*, with causal knowledge, paralleling the knowledge that a structural analyst would encode in equations or in graphs.

For example, to communicate the understanding that in a randomized clinical trial (see Figure 7(b)) the way subjects react (Y) to treatments (X) is statistically independent of the treatment assignment (Z), the analyst would write $Y(x) \perp\!\!\!\perp Z$. Likewise, to convey the understanding that the assignment processes is randomized, hence independent of any variation in the treatment selection process, structurally written $U_X \perp\!\!\!\perp U_Z$, the analyst would use the independence constraint $X(z) \perp\!\!\!\perp Z$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest, for example, $P^*(Y(x) = y)$; in other cases, only bounds on the solution can be obtained. Section 6 explains why this approach is conceptually appealing to some statisticians, even though the process of eliciting judgments about counterfactual dependencies has so far not been systematized. When counterfactual variables are not viewed as by-products of a deeper, process-based model, it is hard to ascertain whether *all* relevant judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent. The elicitation of such judgments can be systematized using the following translation from graphs.

Graphs provide qualitative information about the structure of both the equations in the model and the probability function $P(u)$, the former is encoded as missing arrows, the latter as missing dashed arcs. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model (38). Hence, missing arrows encode exclusion assumptions, that is, claims that adding excluded variables to an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among disturbance terms in two or more equations. For example, the absence of dashed arcs between a node Y and a set of nodes $Z_1, \ldots, Z_k$ implies that the corresponding error variables, $U_Y, U_{Z_1}, \ldots, U_{Z_k}$, are jointly independent in $P(u)$.

These assumptions can be translated into the counterfactual notation using two simple rules; the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. Exclusion restrictions: For every variable Y having parents PA_Y, and for every set of variables S disjoint of PA_Y, we have

$$Y(pa_Y) = Y(pa_Y, s) \tag{41}$$

2. Independence restrictions: If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, we have

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\} \tag{42}$$

Given a sufficient number of such restrictions on P^*, it is possible to compute causal effects $P^*(Y(x) = y)$ using standard probability calculus together with the logical constraints (e.g., Eq. (40) that couple counterfactual variables with their measurable counterparts. These constraints can be used as axioms, or rules of inference, in attempting to transform causal effect expressions, $P^*(Y(x) = y)$, into expressions involving only measurable variables. When such a transformation is found, the corresponding causal effect is identifiable, since P^* reduces then to P. The axioms needed for such transformation are:

$$\textbf{Degeneracy}: \quad Y(\emptyset) = Y \tag{43}$$

$$\textbf{Composition}: \quad Y(x) = Y(x, Z(x)) \text{ for any } Z \text{ disjoint of } \{X, Y\} \tag{44}$$

$$\textbf{Sure}-\textbf{thing}: \quad \text{If } Y(x,z) = Y(x',z)\ \forall\ x' \neq x, \text{ then } Y(x,z) = Y(z) \tag{45}$$

Degeneracy asserts that the observed value of Y is equivalent to a counterfactual variable $Y(x)$ in which the conditional part: "had X been x" is not enforced, that is, X is the empty set.

The Composition axiom asserts:

$$\text{If } Y(x,z) = y \text{ and } Z(x) = z, \text{ then } Y(x) = y$$

and, conversely:

$$\text{If } Y(x) = y \text{ and } Z(x) = z, \text{ then } Y(x,z) = y$$

In words: "The value that Y would obtain had X been x is the same as that obtained had X been x and Z been z, where z is the value that Z would obtain had X been x".

The sure-thing axiom (named after Savage's "sure-thing principle") asserts that if $Y(x,z) = y$ for every value x of X, then the counterfactual antecedant $X = x$ is redundant, namely, we need not concern ourselves with the value that X actually obtains.

Properties (44)-(45) are theorems in the structural interpretation of $Y(x,u)$ as given in Eq. (39). However, in the Neyman-Rubin model, where $Y(x,u)$ is taken as a primitive notion, these properties must be considered axioms which, together with other such properties, defines the abstract counterfactual conditioning operator "had X been x".[7] It is easy to verify that composition and degeneracy imply the consistency rule of (40); substituting $X = \{\emptyset\}$ in (47) yields $Y = Y(z)$ if $Z = z$, which is equivalent to (40).

As an example, let us compute the causal effects associated with the model shown in Figure 5. The parents sets a given by:

$$PA_X = \{\emptyset\},\ PA_Z = \{X\},\ PA_Y = \{Z\} \tag{46}$$

Consequently, the exclusion restrictions translate into:

$$\begin{aligned} Z(x) &= Z(y,x) & (47)\\ X(y) &= X(z,y) = X(z) = X & (48)\\ Y(z) &= Y(z,x) & (49) \end{aligned}$$

The absence of a dashed arc between Z and $\{Y, X\}$ translates into the independence restrictions:

$$Z(x) \perp\!\!\!\perp \{Y(z), X\} \tag{50}$$

Task-1, compute $P^*(Z(x) = z)$ (Equivalently $P(z|\hat{x})$)
From (50) we have $Z(x) \perp\!\!\!\perp X$, hence

$$P^*(Z(x) = z) = P^*(Z(x) = z|x) = P^*(z|x) = P(z|x) \tag{51}$$

[7]The composition rule was communicated to me by James Robins (1995, in conversation) as a property needed for defining graph models (called "finest fully randomized causal graphs" in (Robins pp. 1419-1423, 1986)); in Robins' treatment of counterfactuals, $Y(x,z)$ and $Z(x)$ may not be defined.

Task-2, compute $P^*(Y(z)=y)$ (Equivalently $P^*(y|\hat{z})$)

$$P^*(Y(z)=y) = \sum_x P^*(Y(z)=y|x)P^*(x) \tag{52}$$

From (50) we have

$$Y(z) \perp\!\!\!\perp Z(x)|X \tag{53}$$

hence

$$\begin{aligned} P^*(Y(z)=y|x) &= P^*(Y(z)=y|x, Z(x)=z) && \text{by (52)} \\ &= P^*(Y(z)=y|x,z) && \text{by (40)} \\ &= P^*(y|x,z) && \text{by (40)} \\ &= P(y|x,z) \end{aligned} \tag{54}$$

Substituting (54) in (52), gives

$$P^*(Y(z)=y) = \sum_x P(y|x,z)P(x) \tag{55}$$

which is the celebrated covariate-adjustment formula for causal effect, as in Eq. (9).

Task-3, compute $P^*(Y(x)=y)$ (Equivalently $P(y|\hat{x})$)
For any arbitrary variable Z, we have (by composition)

$$Y(x) = Y(x, Z(x))$$

In particular, since $Y(x,z) = Y(z)$ (from (49)), we have

$$Y(x) = Y(x, Z(x)) = Y(Z(x))$$

and

$$\begin{aligned} P^*(Y(x)=y) &= P^*(Y(Z(x))=y) \\ &= \sum_z P^*(Y(Z(x))=y)|Z(x)=z) \ P^*(Z(x)=z) \\ &= \sum_z P^*(Y(z)=y)|Z(x)=z) \ P^*(Z(x)=z) \\ &= \sum_z P^*(Y(z)=y) \ P^*(Z(x)=z) \end{aligned}$$

since $Y(z) \perp\!\!\!\perp Z(x)$.

$P^*(Y(z)=y)$ and $P^*(Z(x)=z)$ were computed in (55) and (51), respectively, hence

$$P^*(Y(x)=y) = \sum_z P(z|x) \sum_{x'} P(y|z,x')P(z')$$

In summary, the structural and counterfactual frameworks are complementary of each other. Structural analysts can interpret counterfactual sentences as constraints over the solution set of a given system of equations (39) and, conversely, counterfactual analysts can use the constraints (over P^*) given by Eqs. (41) and (42) as a definition of graphs, structural equations and the physical processes which they represent.

ACKNOWLEDGMENT

Much of this investigation was inspired by Spirtes et al. (1993), in which a graphical account of manipulations was first proposed. Phil Dawid, David Freedman, James Robins and Donald Rubin have provided genuine encouragement and valuable advice. For example, Phil has suggested the first part of Definition 2. The investigation also benefitted from discussions with Joshua Angrist, Peter Bentler, David Cox, Arthur Dempster, David Galles, Arthur Goldberger, Sander Greenland, David Hendry, Paul Holland, Guido Imbens, Ed Leamer, Rod McDonald, John Pratt, Paul Rosenbaum, Keunkwan Ryu, Glenn Shafer, Michael Sobel, David Tritchler, and Nanny Wermuth. The research was partially supported by grants from AFOSR and NSF.

REFERENCES

ANGRIST, J.D., IMBENS, G.W., & RUBIN, D.B. (1993). Identification of causal effects using instrumental variables. Department of Economics, Harvard University, Cambridge, MA, Technical Report No. 136. To appear in *JASA*.

BALKE, A., & PEARL, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence*, Eds R. Lopez de Mantaras and D. Poole, 46–54. Morgan Kaufmann: San Mateo, CA.

BOWDEN, R.J., & TURKINGTON, D.A. (1984). *Instrumental Variables*, Cambridge University Press: Cambridge, MA.

COX, D.R. (1958). *The Planning of Experiments*, John Wiley and Sons: New York.

COX, D.R. (1992). Some statistical aspects. *Journal of the Royal Statistical Society, Series A* **155**, 291–301.

COX, D.R., & WERMUTH, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science* **8**, 204–218.

DAWID, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series A* **41**, 1–31.

FISHER, F.M. (1970). A correspondence principle for simultaneous equation models. *Econometrica* **38**, 73–92.

FREEDMAN, D. (1987). As others see us: A case study in path analysis, (with discussion). *Journal of Educational Statistics* **12**, 101–223.

FRISCH, R. (1938). Statistical versus theoretical relations in economic macrodynamics. *League of Nations Memorandum.* (Reproduced in *Autonomy of Economic Relations*, Universitetets Socialokonomiske Institutt, Olso, 1948).

GEIGER, D., VERMA, T.S., & PEARL, J. (1990). Identifying in dependence in Bayesian networks. *Networks* **20**, 507–534.

GOLDBERGER, A.S. (1973). *Structural Equation Models in the Social Sciences*, Seminar Press: New York.

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

HOLLAND, P.W. (1988). Causal inference, path analysis, and recursive structural equations models. in *Sociological Methodology*, Ed C. Clogg, 449–484. American Sociological Association: Washington, DC.

IMBENS, G.W. & ANGRIST, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–476.

LAURITZEN, S.L., & SPIEGELHALTER, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Proceedings of the Royal Statistical Society. Series B* **50**, 154–227.

LAURITZEN, S.L., DAWID, A.P., LARSEN, B.N., & LEIMER, H.G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.

MANSKI, C.F. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* **80**, 319–323.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligence Systems.* Morgan Kaufmann: San Mateo, CA, 1988. Revised 2nd printing, 1992.

PEARL, J. (1993a). Belief networks revisited. *Artificial Intelligence* **59**, 49–56.

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

PEARL, J. (1993c). Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 391–401. Tome LV, Book 1: Florence, Italy,

PEARL, J. (1994a). From bayesian networks to causal networks. In *Bayesian Networks and Probabilistic Reasoning*, Ed A. Gammerman, 1–31. Alfred Walter Ltd.: London.

PEARL, J. (1994b). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence*, Eds R. Lopez de Mantaras and D. Poole, 454–462. Morgan Kaufmann: San Mateo, CA.

PEARL, J. & ROBINS, J. (1995). Causal effects of dynamic policies. In preparation.

PEARL, J. & VERMA, T. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, Eds J.A. Allen, R. Fikes and E. Sandewall, 441–452. Morgan Kaufmann: San Mateo, CA.

PRATT, J.W., & SCHLAIFER, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics* **39**, 23–52.

ROBINS, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modelling* **7**, 1393–1512.

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

ROBINS, J.M., BLEVINS, D., RITTER, G., & WULFSOHN, M. (1992). G-Estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–336.

ROSENBAUM, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* (General) **147**, 656–666.

ROSENBAUM, P., & RUBIN, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrica* **70**, 41–55.

RUBIN, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

RUBIN, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **7**, 34–58.

RUBIN, D.B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.

SIMON, H.A. (1953). Causal Ordering and Identifiability. *Studies in Econometric Method*, Eds W.C. Hood and T.C. Koopmans, Chapter 3. John Wiley and Sons: New York.

SOBEL, M.E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika* **55**, 495–515.

SPIEGELHALTER, D.J., LAURITZEN, S.L., DAWID, P.A., & COWELL, R.G. (1993). Bayesian analysis in expert systems. *Statistical Science* **8**, 219–247.

SPIRTES, P. (1994). Conditional independence in directed cyclic graphical models for feedback. Department of Philosophy, Carnegie-Mellon University, Pittsburg, PA, Technical Report CMU-PHIL-53. To appear.

SPIRTES, P., GLYMOUR, C., & SCHIENES, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag: New York.

STROTZ, R.H., & WOLD, H.O.A. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28**, 417–427.

WAINER, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics* **14**, 121–140.

WERMUTH, N. (1992). On block-recursive regression equations, (with discussion). *Brazilian Journal of Probability and Statistics* **6**, 1–56.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics.* John Wiley and Sons: Chichester, England.

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