Biometrika (1995), **82**, 4, *pp.* 669–710
Printed in Great Britain

TECHNICAL REPORT
R-218-B

Causal diagrams for empirical research (With Discussions)

BY JUDEA PEARL

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, California 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.

Some key words: Causal inference; Graph model; 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 study is complicated by several factors. First, controlled randomised experiments are infeasible: farmers insist on deciding for themselves which plots are to be fumigated. Secondly, farmers' choice of treatment depends on last year's eelworm population, Z_0, an unknown quantity 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. Instead, eelworm 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.

(i) Explicate the assumptions underlying the model.

(ii) Decide whether the assumptions are sufficient for obtaining consistent estimates of the target quantity: the total effect of the fumigants on yields.
(iii) If the answer to (ii) is affirmative, provide a closed-form expression for the target quantity, in terms of distributions of observed quantities.
(iv) If the answer to (ii) is negative, suggest 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 Fig. 1, which represents the investigator's understanding of the major causal influences among measurable quantities in the domain. The quantities Z_1, Z_2 and Z_3 denote, respectively, the eelworm population, both size and type, before treatment, after treatment, and at the end of the season. Quantity Z_0 represents last year's eelworm population; because it is an unknown quantity, it is represented by a hollow circle, as is 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 cannot affect oat plants 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 non-experimental data, for example, estimating the total effect of fumigants on yields.

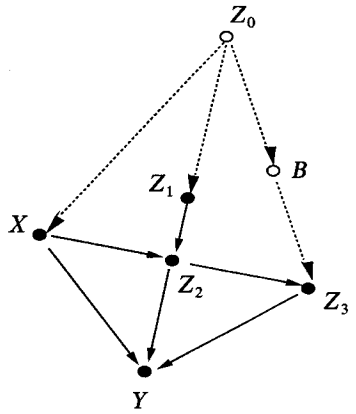

Fig. 1. A causal diagram representing the effect of fumigants, X, on yields, Y.

The proposed method allows an investigator to inspect the diagram of Fig. 1 and conclude immediately the following.

(a) 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.
(b) The total effect of X on Y, assuming discrete variables throughout, is given by the formula

$$\text{pr}(y|\check{x}) = \sum_{z_1}\sum_{z_2}\sum_{z_3} \text{pr}(y|z_2, z_3, x)\,\text{pr}(z_2|z_1, x)\sum_{x'} \text{pr}(z_3|z_1, z_2, x')\,\text{pr}(z_1, x'), \quad (1)$$

where the symbol $\check{x}$, read 'x check', denotes that the treatment is set to level $X = x$ by external intervention.

(c) 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 $\text{pr}(y|\hat{x})$.

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

The formal semantics of the causal diagrams used in this paper will be defined in § 2, following a review of directed acyclic graphs as a language for communicating conditional independence assumptions. Section 2·2 introduces a causal interpretation of directed graphs based on nonparametric structural equations and demonstrates their use in predicting the effect of interventions. Section 3 demonstrates the use of causal diagrams to control confounding bias in observational studies. We 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 § 4, we introduce a symbolic calculus that permits the stepwise derivation of causal effect formulae of the type shown in (1). Using this calculus, § 5 characterises 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 as economical schemes for representing conditional independence assumptions is well evidenced in the literature (Pearl, 1988; Whittaker, 1990). It stems from the existence of graphical methods for identifying the conditional independence relationships implied by recursive product decompositions

$$\text{pr}(x_1, \ldots, x_n) = \prod_i \text{pr}(x_i | pa_i), \tag{2}$$

where pa_i stands for the realisation of some subset of the variables that precede X_i in the order $(X_1, X_2, \ldots, X_n)$. If we construct a directed acyclic graph in which the variables corresponding to 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 graph using the following test.

Definition 1 (*d-separation*). *Let X, Y and Z be three disjoint subsets of nodes in a directed acyclic graph G, and let p be any path between a node in X and a node in Y, where by 'path' we mean any succession of arcs, regardless of their directions. Then 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. Further, Z is said to d-separate X from Y, in G, written $(X \perp\!\!\!\perp Y | Z)_G$, if and only if 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 (2), and the set of triples (X, Z, Y) that satisfy the d-separation criterion in G (Geiger, Verma & Pearl, 1990).

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

2·2. *Graphs as models of interventions*

The use of directed acyclic graphs as carriers of independence assumptions has also been instrumental in predicting the effect of interventions when these graphs are given a causal interpretation (Spirtes, Glymour & Scheines, 1993, p. 78; Pearl, 1993b). Pearl (1993b), for example, treated interventions as variables in an augmented probability space, and their effects were obtained by ordinary conditioning.

In this paper we pursue a different, though equivalent, causal interpretation of directed graphs, based on nonparametric structural equations, 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 Fig. 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 directed graph G represents a deterministic function

$$X_i = f_i(pa_i, \varepsilon_i) \quad (i = 1, \ldots, n), \tag{3}$$

where pa_i denote the parents of variable X_i in G, and ε_i $(1 \leqslant i \leqslant n)$ are mutually independent, arbitrarily distributed random disturbances (Pearl & Verma, 1991). These disturbance terms represent 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 or B in Fig. 1. For example, the causal assumptions conveyed by the model in Fig. 1 correspond to the following set of equations:

$$\begin{gathered} Z_0 = f_0(\varepsilon_0), \quad Z_2 = f_2(X, Z_1, \varepsilon_2), \quad B = f_B(Z_0, \varepsilon_B), \quad Z_3 = f_3(B, Z_2, \varepsilon_3), \\ Z_1 = f_1(Z_0, \varepsilon_1), \quad Y = f_Y(X, Z_2, Z_3, \varepsilon_Y), \quad X = f_X(Z_0, \varepsilon_X). \end{gathered} \tag{4}$$

The equational model (3) is the nonparametric analogue of a structural equations model (Wright, 1921; Goldberger, 1972), 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 such equations convey the asymmetrical counterfactual relation 'is determined by', 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 $(X, Z_2, Z_3, \varepsilon_Y)$ were to assume the values $(x, z_2, z_3, \varepsilon_Y)$, respectively, Y would take on the value dictated by the function f_Y. Thus, the corresponding potential response variable in Rubin's model $Y_{(x)}$, the value that Y would take if X were x, becomes a deterministic function of Z_2, Z_3 and

ε_Y, whose distribution is thus determined by those of Z_2, Z_3 and ε_Y. The relation between graphical and counterfactual models is further analysed by Pearl (1994a).

Characterising each child-parent relationship as a deterministic function, instead of by the usual conditional probability $\text{pr}(x_i|pa_i)$, imposes equivalent independence constraints on the resulting distributions, and leads to the same recursive decomposition (2) that characterises directed acyclic graph models. This occurs because each ε_i is independent of all nondescendants of X_i. However, the functional characterisation $X_i = f_i(pa_i, \varepsilon_i)$ 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 to a selected 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 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 $X_i = f_i(pa_i, \varepsilon_i)$ and placing it under the influence of a new mechanism that sets its value to x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $\text{set}(X_i = x_i)$, or $\text{set}(x_i)$ for short, amounts to removing the equation $X_i = f_i(pa_i, \varepsilon_i)$ from the model, and substituting x_i for X_i in the remaining equations. The model thus created represents the system's behaviour under the intervention $\text{set}(X_i = x_i)$ and, when solved for the distribution of X_j, yields the causal effect of X_i on X_j, denoted by $\text{pr}(x_j|\check{x}_i)$. More generally, when an intervention forces a subset X of variables to fixed values x, a subset of equations is to be pruned from the model given in (3), one for each member of X, thus defining a new distribution over the remaining variables, which completely characterises the effect of the intervention. We thus introduce the following.

DEFINITION 2 (*causal effect*). *Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted* $\text{pr}(y|\check{x})$, *is a function from X to the space of probability distributions on Y. For each realisation x of X,* $\text{pr}(y|\check{x})$ *gives the probability of* $Y = y$ *induced on deleting from the model* (3) *all equations corresponding to variables in X and substituting x for X in the remainder.*

An explicit translation of intervention into 'wiping out' equations from the model was first proposed by Strotz & Wold (1960), and used by Fisher (1970) and Sobel (1990). Graphical ramifications were explicated by Spirtes et al. (1993) and Pearl (1993b). A related mathematical model using event trees has been introduced by Robins (1986, pp. 1422–5).

Regardless of whether we represent interventions as a modification of an existing model as in Definition 2, or as a conditionalisation in an augmented model (Pearl, 1993b), the result is a well-defined transformation between the pre-intervention and the post-intervention distributions. In the case of an atomic intervention $\text{set}(X_i = x_i')$, this transformation can be expressed in a simple algebraic formula that follows immediately from (3) and Definition 2:

$$\text{pr}(x_1, \ldots, x_n|\check{x}_i') = \begin{cases} \{\text{pr}(x_1, \ldots, x_n)\}/\{\text{pr}(x_i|pa_i)\} & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \tag{5}$$

This formula reflects the removal of the terms $\text{pr}(x_i|pa_i)$ from the product in (2), since pa_i no longer influence X_i. Graphically, this is equivalent to removing the links between pa_i

and X_i while keeping the rest of the network intact. Equation (5) can also be obtained from the G-computation formula of Robins (1986, p. 1423) and the Manipulation Theorem of Spirtes et al. (1993), who state that this formula was 'independently conjectured by Fienberg in a seminar in 1991'. Clearly, an intervention set(x_i) can affect only the descendants of X_i in G. Additional properties of the transformation defined in (5) are given by Pearl (1993b).

The immediate implication of (5) is that, given a causal diagram in which all parents of manipulated 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 Fig. 1, where some members of pa_i may be unobservable, thus preventing estimation of $\text{pr}(x_i|pa_i)$. The next two sections provide simple graphical tests for deciding when $\text{pr}(x_j|\check{x}_i)$ 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 $\text{pr}(x_j|\check{x}_i)$ from a sample estimate of $\text{pr}(V_0)$.

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). Pearl (1993b) shows that such judgments are equivalent to a simple graphical test, named the 'back-door criterion', which can be applied directly to the causal diagram.

Definition 3 (*Back-door criterion*). *A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a directed acyclic graph 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. If X and Y are two disjoint sets of nodes in G, Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies it 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 blocked; these paths can be viewed as entering X_i through the back door. In Fig. 2, 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 $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. 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 § 4·4.

We summarise 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* $\text{pr}(y|\check{x})$ *can be computed uniquely from any positive distribution of the observed variables that is compatible with G.*

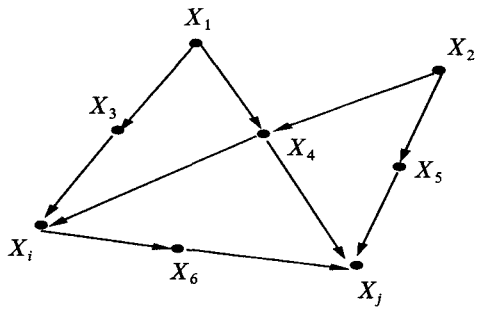

Fig. 2. A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ or $\{X_4, X_5\}$ yields a consistent estimate of $\text{pr}(x_j|\check{x}_i)$.

Identifiability means that $\text{pr}(y|\check{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution. To prove nonidentifiability, it is sufficient to present two sets of structural equations, both complying with (3), 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*

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(y|x, z)\,\text{pr}(z). \tag{6}$$

Equation (6) represents the standard 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, to minimise measurement cost or sampling variability.

3·2. *The front-door criteria*

An alternative criterion, 'the front-door criterion', may be applied in cases where we cannot find observed covariates Z satisfying the back-door conditions. Consider the diagram in Fig. 3. Although Z does not satisfy any of the back-door conditions, measurements of Z nevertheless enable consistent estimation of $\text{pr}(y|\check{x})$. This can be shown by reducing the expression for $\text{pr}(y|\check{x})$ to formulae computable from the observed distribution function $\text{pr}(x, y, z)$.

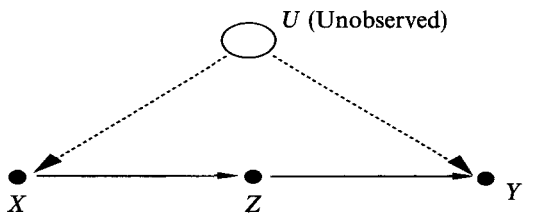

Fig. 3. A diagram representing the front-door criterion.

The joint distribution associated with Fig. 3 can be decomposed into

$$\text{pr}(x, y, z, u) = \text{pr}(u)\,\text{pr}(x|u)\,\text{pr}(z|x)\,\text{pr}(y|z, u) \tag{7}$$

and, from (5), the causal effect of X on Y is given by

$$\text{pr}(y|\check{x}) = \sum_{u} \text{pr}(y|x, u)\,\text{pr}(u). \tag{8}$$

Using the conditional independence assumptions implied by the decomposition (7), we can eliminate u from (8) to obtain

$$\text{pr}(y|\check{x}) = \sum_{z} \text{pr}(z|x) \sum_{x'} \text{pr}(y|x', z)\,\text{pr}(x'). \tag{9}$$

We summarise this result by a theorem.

THEOREM 2. *Suppose a set of variables Z satisfies the following conditions relative to an ordered pair of variables (X, Y): (i) Z intercepts all directed paths from X to Y, (ii) there is no back-door path between X and Z, and (iii) every back-door path between Z and Y is blocked by X. Then the causal effect of X on Y is identifiable and is given by* (9).

The graphical criterion of Theroem 2 uncovers many new structures that permit the identification of causal effects from measurements of variables that are affected by treatments: see § 5·2. The relevance of such structures in practical situations can be seen, for instance, if we identify X with smoking, Y with lung cancer, Z with the amount of tar deposited in a subject's lungs, and U with an unobserved carcinogenic genotype that, according to some, also induces an inborn craving for nicotine. In this case, (9) would provide us with the means to quantify, from nonexperimental data, the causal effect of smoking on cancer, assuming, of course, that $\text{pr}(x, y, z)$ is available and that we believe that smoking does not have any direct effect on lung cancer except that mediated by tar deposits.

4. A CALCULUS OF INTERVENTION

4·1. *General*

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 shall 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 $\text{pr}(y|\check{x})$, where X and Y denote sets of observed variables. By derivation we mean step-wise reduction of the expression $\text{pr}(y|\check{x})$ 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.

4·2. *Preliminary notation*

Let X, Y and Z be arbitrary disjoint sets of nodes in a directed acyclic graph 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

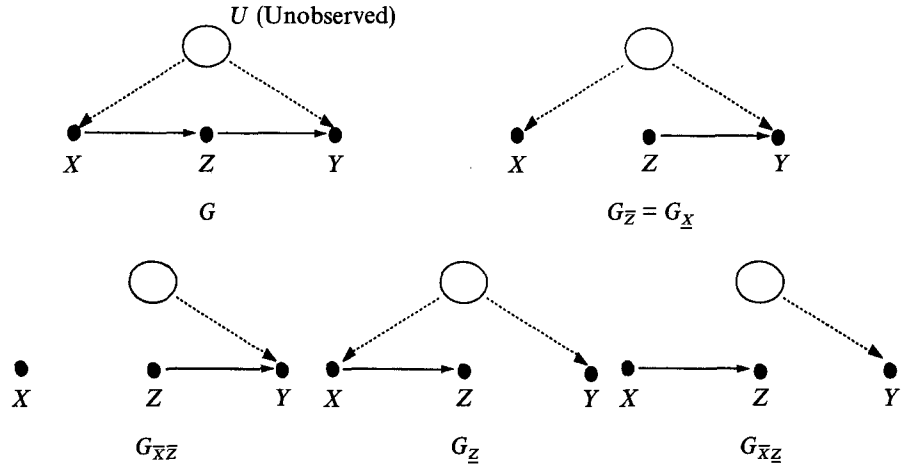

Fig. 4. Subgraphs of G used in the derivation of causal effects.

use the notation $G_{\overline{X}\underline{Z}}$: see Fig. 4 for illustration. Finally, $\text{pr}(y|\check{x}, z) := \text{pr}(y, z|\check{x})/\text{pr}(z|\check{x})$ denotes the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

4·3. *Inference rules*

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

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

Rule 1 (*insertion/deletion of observations*):

$$\text{pr}(y|\check{x}, z, w) = \text{pr}(y|\check{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{10}$$

Rule 2 (*action/observation exchange*):

$$\text{pr}(y|\check{x}, \check{z}, w) = \text{pr}(y|\check{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{11}$$

Rule 3 (*insertion/deletion of actions*):

$$\text{pr}(y|\check{x}, \check{z}, w) = \text{pr}(y|\check{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}, \overline{Z(W)}}}, \tag{12}$$

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 '$\check{x}$' operator as a replacement of the causal mechanism that connects X to its pre-intervention parents by a new mechanism $X = x$ introduced by intervening force. The result is a submodel characterised by the subgraph $G_{\overline{X}}$, called the 'manipulated graph' by 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 (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 $G_{\underline{X}}$, since $G_{\underline{X}\underline{Z}}$ 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.

COROLLARY 1. *A causal effect* $q = \mathrm{pr}(y_1, \ldots, y_k | \check{x}_1, \ldots, \check{x}_m)$ *is identifiable in a model characterised 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. check-free, probability expression involving observed quantities.*

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 systematised and executed by efficient algorithms as described by Galles & Pearl (1995). As § 4·4 illustrates, symbolic derivations using the check notation are much more convenient than algebraic derivations that aim at eliminating latent variables from standard probability expressions, as in § 3·2.

4·4. *Symbolic derivation of causal effects: An example*

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

Task 1: *compute* $\mathrm{pr}(z|\check{x})$. 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

$$\mathrm{pr}(z|\check{x}) = \mathrm{pr}(z|x). \tag{13}$$

Task 2: *compute* $\mathrm{pr}(y|\check{z})$. Here we cannot apply Rule 2 to exchange $\check{z}$ 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:

$$\mathrm{pr}(y|\check{z}) = \sum_x \mathrm{pr}(y|x, \check{z})\,\mathrm{pr}(x|\check{z}). \tag{14}$$

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

$$\mathrm{pr}(x|\check{z}) = \mathrm{pr}(x) \quad \text{if } (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}}, \tag{15}$$

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 $\mathrm{pr}(y|x, \check{z})$, we consult Rule 2:

$$\mathrm{pr}(y|x, \check{z}) = \mathrm{pr}(y|x, z) \quad \text{if } (Z \perp\!\!\!\perp Y | X)_{G_{\underline{Z}}}, \tag{16}$$

noting that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write (14) as

$$\mathrm{pr}(y|\check{z}) = \sum_x \mathrm{pr}(y|x, z)\,\mathrm{pr}(x) = E_x\,\mathrm{pr}(y|x, z), \tag{17}$$

which is a special case of the back-door formula (6). The legitimising 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* $\text{pr}(y|\check{x})$. Writing

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(y|z, \check{x})\,\text{pr}(z|\check{x}), \tag{18}$$

we see that the term $\text{pr}(z|\check{x})$ was reduced in (13) but that no rule can be applied to eliminate the 'check' symbol from the term $\text{pr}(y|z, \check{x})$. However, we can add a 'check' symbol to this term via Rule 2:

$$\text{pr}(y|z, \check{x}) = \text{pr}(y|\check{z}, \check{x}), \tag{19}$$

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

$$\text{pr}(y|z, \check{x}) = \text{pr}(y|\check{z}), \tag{20}$$

which was calculated in (17). Substituting (17), (20) and (13) back into (18) finally yields

$$\text{pr}(y|\check{x}) = \sum_z \text{pr}(z|x) \sum_{x'} \text{pr}(y|x', z)\,\text{pr}(x'), \tag{21}$$

which is identical to the front-door formula (9).

The reader may verify that all other causal effects, for example, $\text{pr}(y, z|\check{x})$ and $\text{pr}(x, z|\check{y})$, can likewise be derived through the rules of Theorem 3. Note that in all the derivations the graph G provides both the license for applying the inference rules and the guidance for choosing the right rule to apply.

4·5. *Causal inference by surrogate experiments*

Suppose we wish to learn the causal effect of X on Y when $\text{pr}(y|\check{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomised experiment. The question arises whether $\text{pr}(y|\check{x})$ can be identified by randomising 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 $\text{pr}(y|\check{x})$ into expressions in which only members of Z carry the check 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) $\text{pr}(y|\check{x})$ is identifiable in $G_{\underline{Z}}$. Indeed, if condition (i) holds, we can write $\text{pr}(y|\check{x}) = \text{pr}(y|\check{x}, \check{z})$, because $(Y \perp\!\!\!\perp Z | X)_{G_{\overline{X}\underline{Z}}}$. But $\text{pr}(y|\check{x}, \check{z})$ stands for the causal effect of X on Y in a model governed by $G_{\underline{Z}}$ which, by condition (ii), is identifiable. Figures 7(e) and 7(h) below illustrate models in which both conditions hold. Translated to our cholesterol example, these conditions 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.

5. GRAPHICAL TESTS OF IDENTIFIABILITY

5·1. *General*

Figure 5 shows simple diagrams in which $\text{pr}(y|\hat{x})$ cannot be identified due to the presence of a bow pattern, i.e. a confounding arc, shown 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 Fig. 1 can be represented as a confounding arc between X and Z_3. A bow-pattern represents an equation $Y = f_Y(X, U, \varepsilon_Y)$, 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 $\text{pr}(y|\hat{x})$ even when it is found in the context of a larger graph, as in Fig. 5(b). This is in contrast to linear models, where the addition of an arc to a bow-pattern can render $\text{pr}(y|\hat{x})$ 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 $b = \partial E(Y|\hat{x})/\partial x$ is not identifiable. However, adding an arc $Z \to X$ 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, p. 12; Angrist, Imbens & Rubin, 1995):

$$b := \frac{\partial}{\partial x} E(Y|\hat{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{R_{yz}}{R_{xz}}. \tag{22}$$

In nonparametric models, adding an instrumental variable Z to a bow-pattern, see Fig. 5(b), does not permit the identification of $\text{pr}(y|\hat{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment, Z, is randomised, hence no link enters Z, but compliance is imperfect. The confounding arc between X and Y in Fig. 5(b) represents unmeasurable factors which influence both subjects' choice of treatment, X, and response to treatment, Y. In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $\text{pr}(y|\hat{x})$ without making additional assumptions on the nature of the interactions between compliance and response (Imbens & Angrist, 1994), as is done, for example, in the approach to instrumental variables developed by Angrist et al. (1995). While the added arc $Z \to X$ permits us to calculate bounds on $\text{pr}(y|\hat{x})$ (Robins, 1989, § 1g; Manski, 1990), and while the upper and lower bounds may even coincide for

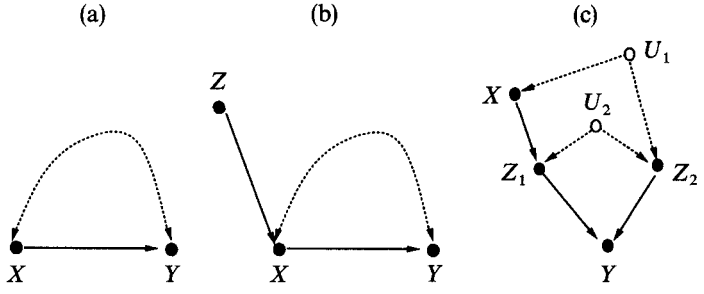

Fig. 5. (a) A bow-pattern: a confounding arc embracing a causal link $X \to Y$, thus preventing the identification of $\text{pr}(y|\hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bow-less graph still prohibiting the identification of $\text{pr}(y|\hat{x})$.

certain types of distributions pr(x, y, z) (Balke & Pearl, 1994), there is no way of computing pr($y|\hat{x}$) for every positive distribution pr(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.

Our ability to compute pr($y|\hat{x}$) for pairs (x, y) of singleton variables does not ensure our ability to compute joint distributions, such as pr(y_1, $y_2|\hat{x}$). Figure 5(c), for example, shows a causal diagram where both pr($z_1|\hat{x}$) and pr($z_2|\hat{x}$) are computable, but pr(z_1, $z_2|\hat{x}$) is not. Consequently, we cannot compute pr($y|\hat{x}$). This diagram is the smallest graph that does not contain a bow-pattern and still presents an uncomputable causal effect.

5·2. *Identifying models*

Figure 6 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 to permit the identification of the target quantity pr($y|\hat{x}$). Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs, shown 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 (3) and then constructing a new diagram by

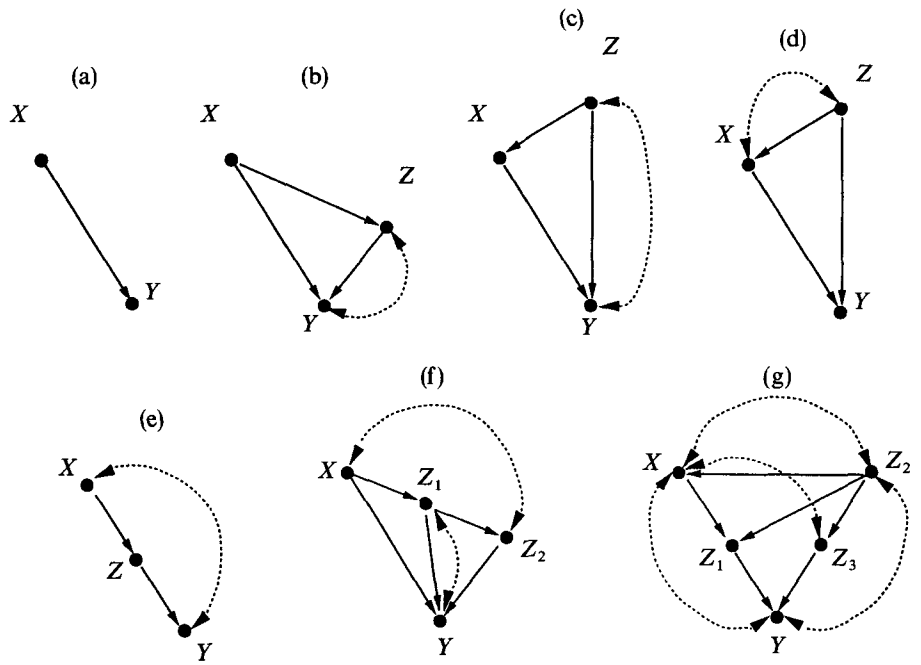

Fig. 6. Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

connecting any two variables X_i and X_j by (i) an arrow from X_j to X_i whenever X_j appears in the equation for X_i, and (ii) 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 Fig. 6.

(i) Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $\text{pr}(y|\hat{x})$ will still be identified in any edge-subgraph of the diagrams shown in Fig. 6. 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, $\text{pr}(y|\hat{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Fig. 6.

(ii) The diagrams in Fig. 6 are maximal, in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $\text{pr}(y|\hat{x})$ no longer identifiable.

(iii) Although most of the diagrams in Fig. 6 contain bow-patterns, none of these patterns emanates from X as is the case in Fig. 7(a) and (b) below. In general, a necessary condition for the identifiability of $\text{pr}(y|\hat{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y.

(iv) Figures 6(a) and (b) 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; that is, X is strongly ignorable relative to Y (Rosenbaum & Rubin, 1983); hence, $\text{pr}(y|\hat{x}) = \text{pr}(y|x)$. Likewise, Figs 6(c) and (d) represent designs in which observed covariates, Z, block every back-door path between X and Y; that is X is conditionally ignorable given Z (Rosenbaum & Rubin, 1983); hence, $\text{pr}(y|\hat{x})$ is obtained by standard adjustment for Z, as in (6):

$$\text{pr}(y|\hat{x}) = \sum_z \text{pr}(y|x, z)\,\text{pr}(z).$$

(v) For each of the diagrams in Fig. 6, we can readily obtain a formula for $\text{pr}(y|\hat{x})$, using symbolic derivations patterned after those in § 4·4. The derivation is often guided by the graph topology. For example, Fig. 6(f) dictates the following derivation. Writing

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

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of Fig. 6(e), with Z_1, Z_2 replacing Z, Y, respectively. Thus, $\text{pr}(z_1, z_2|\hat{x})$ can be obtained from (14) and (21). Likewise, the term $\text{pr}(y|z_1, z_2, \hat{x})$ can be reduced to $\text{pr}(y|z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\underline{X}}}$. Thus, we have

$$\text{pr}(y|\hat{x}) = \sum_{z_1,z_2} \text{pr}(y|z_1, z_2, x)\,\text{pr}(z_1|x) \sum_{x'} \text{pr}(z_2|z_1, x')\,\text{pr}(x'). \tag{23}$$

Applying a similar derivation to Fig. 6(g) yields

$$\text{pr}(y|\hat{x}) = \sum_{z_1} \sum_{z_2} \sum_{x'} \text{pr}(y|z_1, z_2, x')\,\text{pr}(x')\,\text{pr}(z_1|z_2, x)\,\text{pr}(z_2). \tag{24}$$

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.

(vi) In Figs 6(e), (f) and (g), the identifiability of $\text{pr}(y|\hat{x})$ is rendered feasible through observed covariates, Z, that are affected by the treatment X, that is 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, p. 48; 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 reasons 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. Figures 6(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, for example 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 (6): see (9), (23) and (24).

(vii) In Figs 6(b), (c) and (f), 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.

5·3. *Nonidentifying models*

Figure 7 presents typical diagrams in which the total effect of X on Y, $\text{pr}(y|\hat{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

(i) All graphs in Fig. 7 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. It is not a sufficient test, though, as is demonstrated by Fig. 6(e), in which the back-door path (dashed) is unblockable, yet $\text{pr}(y|\hat{x})$ is identifiable.

(ii) A sufficient condition for the nonidentifiability of $\text{pr}(y|\hat{x})$ is the existence of a confounding path between X and any of its children on a path from X to Y, as shown in Figs 7(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Fig. 7 as an edge-subgraph.

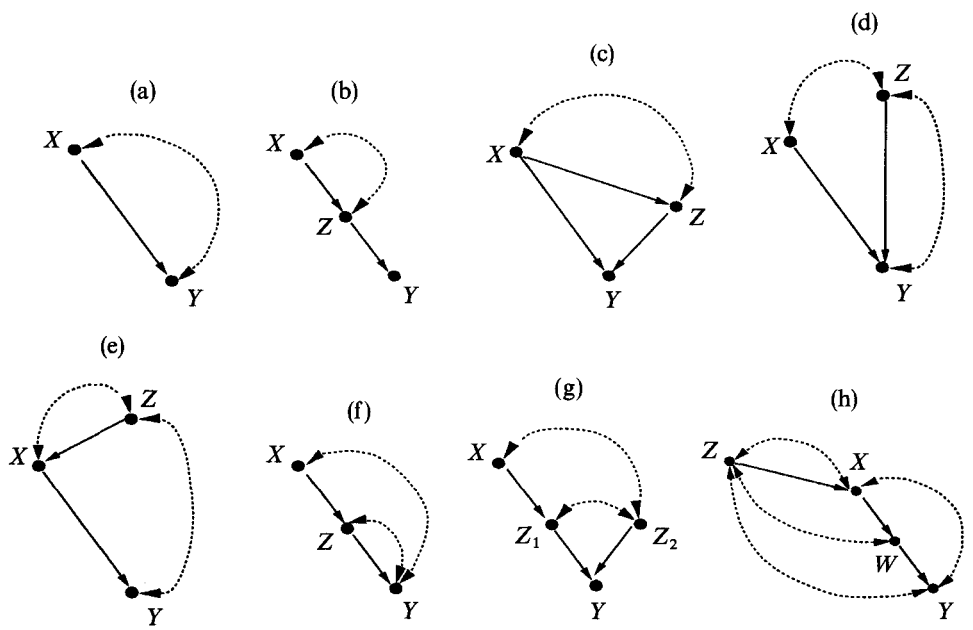

Fig. 7. Typical models in which $\text{pr}(y|\hat{x})$ is not identifiable.

(iii) Figure 7(g) demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $\text{pr}(z_1|\check{x})$, $\text{pr}(z_2|\check{x})$, $\text{pr}(y|\check{z}_1)$ and $\text{pr}(y|\check{z}_2)$, but not $\text{pr}(y|\check{x})$. 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 related papers (Pearl, 1994a, 1995) we show that some of the assumptions, most notably those associated with instrumental variables, see Fig. 5(b), are subject to falsification tests. Additionally, 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, 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 estimates of these with confidence intervals and significance levels, as in traditional analysis of controlled experiments. Having nonparametric estimates 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 linearity are deemed reasonable, then the estimand in (9) can be replaced by $E(Y|\check{x}) = R_{xz}\beta_{zy\cdot x}x$, where $\beta_{zy\cdot x}$ is the standardised regression coefficient, and the estimation problem reduces to that of estimating coefficients. More sophisticated estimation techniques are given by Rubin (1978), Robins (1989, § 17), and Robins et al. (1992, pp. 331–3).

Several extensions of the methods proposed in this paper are possible. First, the analysis of atomic interventions can be generalised 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 $P^*(x|z)$. Pearl (1994b) shows that computing the effect of such policies is equivalent to computing the expression $\text{pr}(y|\check{x}, z)$.

A second extension concerns the use of the intervention calculus of Theorem 3 in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in terms 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). Secondly, the d-separation criterion for directed acyclic graphs 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 (Spirtes, 1995). However, the computation of causal effect estimands will be harder in cyclic networks, because symbolic reduction of $\text{pr}(y|\check{x})$ to check-free expressions may require the solution of nonlinear equations.

Finally, a few comments regarding the notation introduced in this paper. There have been three approaches to expressing causal assumptions in mathematical form. The most

common approach in the statistical literature invokes Rubin's model (Rubin, 1974), in which probability functions are defined over an augmented space of observable and counterfactual variables. In this model, causal assumptions are expressed as independence constraints over the augmented probability function, as exemplified by Rosenbaum & Rubin's (1983) definitions of ignorability conditions. An alternative but related approach, still using the standard language of probability, is to define augmented probability functions over variables representing hypothetical interventions (Pearl, 1993b).

The language of structural models, which includes path diagrams (Wright, 1921) and structural equations (Goldberger, 1972) represents a drastic departure from these two approaches, because it invokes new primitives, such as arrows, disturbance terms, or plain causal statements, which have no parallels in the language of probability. This language has been very popular in the social sciences and econometrics, because it closely echoes statements made in ordinary scientific discourse and thus provides a natural way for scientists to communicate knowledge and experience, especially in situations involving many variables.

Statisticians, however, have generally found structural models suspect, because the empirical content of basic notions in these models appears to escape conventional methods of explication. For example, analysts have found it hard to conceive of experiments, however hypothetical, whose outcomes would be constrained by a given structural equation. Standard probability calculus cannot express the empirical content of the coefficient b in the structural equation $Y = bX + \varepsilon_Y$ 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 this equation certain variables that are highly correlated with X or Y. As a consequence, the whole enterprise of structural equation modelling has become the object of serious controversy and misunderstanding among researchers (Freedman, 1987; Wermuth, 1992; Whittaker, 1990, p. 302; Cox & Wermuth, 1993).

To a large extent, this history of controversy stems not from faults in the structural modelling approach but rather from a basic limitation of standard probability theory: when viewed as a mathematical language, it is too weak to describe the precise experimental conditions that prevail in a given study. For example, standard probabilistic notation cannot distinguish between an experiment in which variable X is observed to take on value x and one in which variable X is set to value x by some external control. The need for this distinction was recognised by several researchers, most notably Pratt & Schlaifer (1988) and Cox (1992), but has not led to a more refined and manageable mathematical notation capable of reflecting this distinction.

The 'check' notation developed in this paper permits one to specify precisely what is being held constant and what is merely measured in a given study and, using this specification, the basic notions of structural models can be given clear empirical interpretation. For example, the meaning of b in the equation $Y = bX + \varepsilon_Y$ is simply $\partial E(Y|\check{x})/\partial x$, 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, for example, via another equation: $X = aY + \varepsilon_X$. Moreover, the notion of randomisation need not be invoked. Likewise, the analyst's decision as to which variables should be included in a given equation is 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, $\text{pr}(y|\check{z}, \check{s}_{YZ}) = \text{pr}(y|\check{s}_{YZ})$. In other words, variables that are excluded from the equation $Y = bX + \varepsilon_Y$ are not conditionally independent of Y given measurements of X, but rather conditionally independent of 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 $Y - E(Y|\check{s}_Y)$; two disturbance terms, ε_X and ε_Y, are correlated if $\text{pr}(y|\check{x}, \check{s}_{XY}) \neq \text{pr}(y|x, \check{s}_{XY})$; and so on.

The distinctions provided by the 'check' notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers. Moreover, since most scientific knowledge is organised 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 when combined with statistical data.

ACKNOWLEDGEMENT

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. 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 Air Force Office of Scientific Research and National Science Foundation.

APPENDIX

Proof of Theorem 3

(i) Rule 1 follows from the fact that deleting equations from the model in (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 characterising this submodel is given by $G_{\overline{X}}$, $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$ implies the conditional independence $\text{pr}(y|\check{x}, z, w) = \text{pr}(y|\check{x}, w)$ in the post-intervention distribution.

(ii) 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 for the functions that determine Z in the structural equations (Pearl, 1993b). 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 $F_Z = \text{set}(z)$.

(iii) The following argument was developed by 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 $\text{pr}(y|\check{x}, \check{z}, w) = \text{pr}(y|\check{x}, w)$. 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 an arrow pointing to Z'', or in 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 $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$ implies $(F_Z \perp\!\!\!\perp Y \mid W, X)_{G_{\overline{X}}}$, and thus $\text{pr}(y \mid \check{x}, \check{z}, w) = \text{pr}(y \mid \check{x}, w)$.

References

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1995). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* To appear.

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

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

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

Cox, D. R. (1992). Causality: Some statistical aspects. *J. R. Statist. Soc.* A **155**, 291–301.

Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statist. Sci.* **8**, 204–18.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc.* B **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). *J. Educ. Statist.* **12**, 101–223.

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

Galles, D. & Pearl, J. (1995). Testing identifiability of causal effects. In *Uncertainty in Artificial Intelligence—11*, Ed. P. Besnard and S. Hanks, pp. 185–95. San Francisco, CA: Morgan Kaufmann.

Geiger, D., Verma, T. S. & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–34.

Goldberger, A. S. (1972). Structural equation models in the social sciences. *Econometrica* **40**, 979–1001.

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, pp. 449–84. Washington, D.C.: American Sociological Association.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–76.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems (with Discussion). *J. R. Statist. Soc.* B **50**, 157–224.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev., Papers Proc.* **80**, 319–23.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1993a). Belief networks revisited. *Artif. Intel.* **59**, 49–56.

Pearl, J. (1993b). Comment: Graphical models, causality, and intervention. *Statist. Sci.* **8**, 266–9.

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

Pearl, J. (1994b). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence*, Ed. R. Lopez de Mantaras and D. Poole, pp. 452–62. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal inference from indirect experiments. *Artif. Intel. Med. J.*, 7, 561–582, 1995.

Pearl, J. & Verma, T. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, Ed. J. A. Allen, R. Fikes and E. Sandewall, pp. 441–52. San Mateo, CA: Morgan Kaufmann.

PRATT, J. W. & SCHLAIFER, R. (1988). On the interpretation and observation of laws. *J. Economet.* **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. *Math. Model.* **7**, 1393–512.
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*, Ed. L. Sechrest, H. Freeman and A. Mulley, pp. 113–59. Washington, D.C.: 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**, 319–36.
ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc.* A **147**, 656–66.
ROSENBAUM, P. & RUBIN, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **7**, 34–58.
RUBIN, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* **5**, 472–80.
SIMON, H. A. (1953). Causal ordering and identifiability. In *Studies in Econometric Method*, Ed. W. C. Hood and T. C. Hoopmans, Ch. 3. New York: John Wiley.
SOBEL, M. E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika* **55**, 495–515.
SPIEGELHALTER, D. J., LAURITZEN, S. L., DAWID, A. P. & COWELL, R. G. (1993). Bayesian analysis in expert systems (with Discussion). *Statist. Sci.* **8**, 219–47.
SPIRTES, P. (1995). Conditional independence in directed cyclic graphical models for feedback. *Networks.* To appear.
SPIRTES, P., GLYMOUR, C. & SCHEINES, R. (1993). *Causation, Prediction, and Search.* New York: Springer-Verlag.
STROTZ, R. H. & WOLD, H. O. A. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28**, 417–27.
WAINER, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *J. Educ. Statist.* **14**, 121–40.
WERMUTH, N. (1992). On block-recursive regression equations (with Discussion). *Brazilian J. Prob. Statist.* **6**, 1–56.
WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: John Wiley.
WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20**, 557–85.

[*Received May* 1994. *Revised February* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY D. R. COX
Nuffield College, Oxford, OX1 1NF, U.K.
AND NANNY WERMUTH
Psychologisches Institut, Johannes Gutenberg-Universität Mainz, Staudingerweg 9, D-55099 Mainz, Germany

Judea Pearl has provided a general formulation for uncovering, under very explicit assumptions, what he calls the causal effect on y of 'setting' a variable x at a specified level, $\mathrm{pr}(y|\check{x})$, as assessed in a system of dependencies that can be represented by a directed acyclic graph. His Theorem 3 then provides a powerful computational scheme.

The back-door criterion requires there to be no unobserved 'common cause' for x and y that is not blocked out by observed variables, that is at least one of the intermediate variables between x and y or the common cause is to be observed. It is precisely doubt about such assumptions that makes epidemiologists, for example, wisely in our view, so cautious in distinguishing risk factors from causal effects. The front-door criterion requires, first, that there be an observed variable z such

that x affects y only via z. Moreover, an unobserved variable u affecting both x and y must have no direct effect on z. Situations where this could be assumed with any confidence seem likely to be exceptional.

We agree with Pearl that in interpreting a regression coefficient, or generalisation thereof, in terms of the effect on y of an intervention on x, it is crucial to specify what happens to other variables, observed and unobserved. Which are fixed, which vary essentially as in the data under analysis, which vary in some other way? If we 'set' diastolic blood pressure, presumably we must, at least for some purposes, also 'set' systolic blood pressure; and what about a host of biochemical variables whose causal interrelation with blood pressure is unclear? The difficulties here are related to those of interpreting structural equations with random terms, difficulties emphasised by Haavelmo many years ago; we cannot see that Pearl's discussion resolves the matter.

The requirement in the standard discussion of experimental design that concomitant variables be measured before randomisation applies to their use for improving precision and detecting interaction. The use of covariates for detailed exploration of the relation between treatment effects, intermediate responses and final responses gets less attention than it deserves in the literature on design of experiments; see, however, the searching discussion in an agronomic context by Fairfield Smith (1957). Graphical models and their consequences have much to offer here and we welcome Dr Pearl's contribution on that account.

[*Received May* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY A. P. DAWID

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.

The clarity which Pearl's graphical account brings to the problems of describing and manipulating causal models is greatly to be welcomed. One point which deserves emphasis is the equivalence, for the purposes Pearl addresses, between the counterfactual functional representation (3), emphasised here, and the alternative formulation of Pearl (1993b), involving the incorporation into a 'regular' directed acyclic graph of additional nodes and links directly representing interventions. I must confess to a strong preference for the latter approach, which in any case is the natural framework for analysis, as is seen from the Appendix. In particular, although a counterfactual interpretation is possible, it is inessential: the important point is to represent clearly, by choice of the appropriate directed acyclic graph, the way in which an intervention set$(X = x)$ disturbs the system, by specifying which conditional distributions are invariant under such an intervention. As (5) makes evident, the overall effect of intervention is then entirely determined by the conditional distributions describing the recursive structure, and in no way depends on the way in which these might be represented functionally as in (3). This is fortunate, since it is far easier to estimate conditional distributions than functional relationships.

There are contexts where distributions are not enough, and counterfactual relationships need to be assessed for valid inference. Perhaps the extension to nonrecursive models mentioned in § 6 is one. More important is inquiry into the 'causes of effects', rather than the 'effects of causes' considered here. This arises in questions of legal liability: 'Did Mr A's exposure to radiation in his workplace cause his child's leukaemia?' Knowing that Mr A was exposed, and the child has developed leukaemia, the question requires us to assess, counterfactually, what would have happened to the child had Mr A not been exposed. For this, distributional models are insufficient: a functional or counterfactual model is essential.

This raises the question as to how we can use scientific understanding and empirical data to construct the requisite causal model. By saying little about this specification problem, Pearl is in danger of being misunderstood to say that it is not important. To build either a distributional or a counterfactual causal model, we need to assess evidence on how interventions affect the system, and what remains unchanged. This will typically require a major scientific undertaking. Given this structure, distributional aspects can, in principle, be estimated from suitable empirical data, if only these are available, and we can then apply the manipulations described by Pearl to address problems of the 'effects of causes'. But much more would be needed to address 'causes of effects', since counterfactual probabilities are, almost by definition, inaccessible to direct empirical study. Empirical data can be used to place bounds on these (Balke & Pearl, 1994), but these will usually only be useful when they essentially determine the functions in (3). And, for this, it will be necessary to conduct studies in which the variables ε_i are explicitly identified and observed. Thus the whole mechanism needs to be broken down into essentially deterministic sub-mechanisms, with randomness arising solely from incomplete observation. In most branches of science such a goal is quite unattainable.

I emphasise the distinction drawn above, between inference about 'effects of causes' and 'causes of effects', because it might be tempting to try to extend Pearl's analysis, particularly in its formulation (3), to the latter problem. For both problems serious difficulties attend the initial model specification, but these are many orders of magnitude greater for 'causes of effects', and the inferences drawn will be very highly sensitive to the specification.

On a different point, I am intrigued by possible connexions between Pearl's clear distinction between conditioning and intervening, and the prequential framework of Dawid (1984, 1991), especially as elaborated by Vovk (1993). Suppose A plays a series of games, involving coins, dice, roulette wheels, etc. At any point, the game chosen may depend on the observed history. We could model this dependence probabilistically, or leave it unspecified. Now suppose we are informed of the sequence of games actually played, and want to say something about their outcomes. In a fully probabilised model, we could condition on the games played, but this would involve unpleasant analysis, and be sensitive to assumptions. Alternatively, and seemingly very reasonably, we can use the 'prequential model', which treats the games as having been fixed in advance. This is obtained from a fully specified model, with its natural temporally defined causal model, by 'setting' the games, rather than conditioning on them.

[*Received May* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY STEPHEN E. FIENBERG

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, U.S.A.

CLARK GLYMOUR AND PETER SPIRTES

Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, U.S.A.

In recent years we have investigated the use of directed graphical models (Spirtes, 1995; Spirtes, Glymour & Scheines, 1993) in order to analyse predictions about interventions that follow from causal hypotheses. We therefore welcome Pearl's development and exposition. Our goal here is to indicate some other virtues of the directed graph approach, and compare it to alternative formalisations.

Directed graph models have a dual role, explicitly representing substantive hypotheses about influence and implicitly representing hypotheses about conditional independence. We can connect the two dimensions, one causal and the other stochastic, by explicit mathematical axioms. For example, the causal Markov axiom requires that, in the graph, each variable be independent of its nondescendants conditional on its set of parents. The formalism allows one to hold causal hypotheses fixed while varying the axiomatic connexions to probabilistic constraints. In this way, one can prove the correctness of computable conditions for prediction, for the statistical equivalence of models, and for the possibility or impossibility of asymptotically correct model search, all under alternative axioms and under a variety of circumstances relevant to causal inference, including the presence of latent variables, sample selection bias, mixtures of causal structures, feedback, etc. Thus it is possible to derive Pearl's Theorem 3, and other results in his paper, from the Markov condition alone, provided one treats a manipulation as conditionalisation on a 'policy' variable appropriately related to the variable manipulated. Further, two extensions of Theorem 3 follow fairly directly. First, if the sufficient conditions in Theorem 3 for the equalities of probabilities are violated, distributions satisfying the Markov condition exist for which the equalities do not hold. Secondly, if the Markov condition entails all conditional independencies holding in a distribution, an axiom sometimes called 'faithfulness', the conditions of Theorem 3 are also necessary for the equalities given there.

The graphical formalism captures many of the essential features common to statistical models that sometimes accompany causal or constitutive hypotheses, including linear and nonlinear regression, factor analysis, and both recursive and nonrecursive structural equation models. In many cases, these models are representable as graphical models with additional distribution assumptions. In some cases, the graphical formalism provides an alternative parametrisation of subsets of the distributions associated with a family of models, as, for example, for the graphical subset of distributions from the log-linear parametrisation of the multinomial family (Bishop, Fienberg & Holland, 1975; Whittaker, 1990). Directed graphs also offer an explicit representation of the connexion between causal hypotheses and independence and conditional independence hypotheses in experimental design, and, under various axioms, permit the mathematical investigation of relations between experimental and nonexperimental designs.

Rubin (1974), Rosenbaum & Rubin (1983), Holland (1988) and Pratt & Schlaifer (1988) have provided an important alternative treatment of the prediction of the results of interventions from partial causal knowledge. As Pearl notes, their approach, which involves conditional independence of measured and 'counterfactual' variables, gives results in agreement with the directed graphical approach under an assumption they refer to as 'strong ignorability'. For example, a result given without proof by Pratt & Schlaifer provides a 'sufficient and almost necessary' condition for the equality of the probability of Y when X is manipulated, and the conditional probability of the counterfactual of Y on X. A direct analogue of their claim of sufficiency is provable from the Markov condition and necessity follows from the faithfulness condition, which is true with probability 1 for natural measures on linear and multinomial parameters. This offers a reasonable reconstruction of what they may have meant by 'almost necessary'. The Rubin approach to prediction has some advantages over directed graph approaches, for example in the representation of circumstances in which features of units influence other units. The disadvantages of the framework stem from the necessity of formulating hypotheses explicitly in terms of the conditional independence of actual and counterfactual variables rather than in terms of variables directly influencing others. In our experience, even experts have difficulty reliably judging the conditional independence relations that do or do not follow from assumptions. For example, we have heard many statistically trained people deny, before doing the calculation, that the normality and independence of X, Y and e, coupled with the linear equation $Z = aX + bY + e$, entail that X and Y are dependent conditional on Z. For the same reason, the Rubin framework may make more difficult mathematical proofs of results about invariance, equivalence, search, etc.

There are at least two other alternative approaches to the graphical formalism: Robins' (1986)

G-computation algorithm for calculating the effects of interventions under causal hypotheses expressed as event trees, an extension of the Rubin approach; and Glenn Shafer's (1996) more recent and somewhat different tree structure approach. Where both are applicable, they seem to give the same results as do procedures Pearl describes for computing on directed graphs. An advantage of the directed graph formalism is the naturalness of the representation of influence. Questions regarding the relative power of these alternative approaches are as follows.

(i) Is the graphical approach applicable to cases where the alternatives are not, particularly when there are structures in which it is not assumed that every variable either influences or is influenced by every other?

(ii) Is the graphical approach faster in some instances, because the directed graphs can encode independencies in their structure while event trees cannot?

(iii) Can the alternatives, like the graphical procedure, be extended to cases in which the distribution forced on the manipulated variable is continuous?

As far as we can tell, none of the approaches to date has been able to cope with causal language associated with explanatory variables in proportional hazards models, where the nonlinear structure does not lend itself naturally to conditional independence representations.

[*Received April* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY DAVID FREEDMAN

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.

Causal inference with nonexperimental data seems unjustifiable to many statisticians. For others, the trick can be done almost on a routine basis, with the help of regression and its allied techniques, like path analysis or simultaneous-equation models. However, typical regression studies are problematic, because inferences are conditional on unvalidated, even unarticulated, assumptions: for discussion and reviews of the literature, see Freedman (1991, 1995).

Deriving causation from association by regression depends on stochastic assumptions of the familiar kind, and on less familiar causal assumptions. Building on earlier work by Holland (1988) and Robins (1989) among others, Pearl develops a graphical language in which the causal assumptions are relatively easy to state. His formulation is both natural and interesting. It captures reasonably well one intuition behind regression analysis: causal inferences can be drawn from associational data if you are observing the results of a controlled experiment run by Nature, and the causal ordering of the variables is known. When these assumptions hold, there is identifiability theory that gives an intriguing description of permissible inferences.

Following Holland (1988), I state the causal assumptions along with statistical assumptions that, taken together, justify inference in conventional path models. There is an observational study with n subjects, $i = 1, \ldots, n$. The data will be analysed by regression. There are three measured variables, X, Y, Z. The path diagram has arrows from X to Y; then, from X and Y to Z. The diagram is interpreted as a set of assumptions about causal structure: the data result from coupling together two thought experiments, as specified below. Statistical analysis proceeds from the assumption that subjects are independent and identically distributed in certain respects. That is the basis for estimating regression functions, an issue Pearl does not address; customary tests of significance would follow too.

Random variables are represented in the usual way on a sample space Ω. With notation like Holland's, $Y_{i,x}(\omega)$ represents the Y-value for subject i at $\omega \in \Omega$, if you set the X-value to x. The

thought experiments are governed by the following assumptions (1) and (2):

$$Y_{i,x}(\omega) = f(x) + \delta_i(\omega), \tag{1}$$

$$Z_{i,x,y}(\omega) = g(x, y) + \varepsilon_i(\omega). \tag{2}$$

The same f and g apply to all subjects. Additive disturbance terms help the regression functions f and g to be estimable, but more is required. Typically, linearity is assumed:

$$f(x) = a + bx, \quad g(x, y) = c + dx + ey. \tag{3}$$

The δ's are taken to be independent and identically distributed with mean 0 and finite variance, as are the ε's; furthermore, the δ's are taken as independent of the ε's.

The experiments are coupled together to produce observables, as follows. Nature assigns X-values to the subjects at random, independently of the δ's and ε's. Finally, the data on subject i are modelled as

$$X_i(\omega), \quad Y_i(\omega) = f\{X_i(\omega)\} + \delta_i(\omega), \quad Z_i(\omega) = g\{X_i(\omega), Y_i(\omega)\} + \varepsilon_i(\omega).$$

Linearity of regression functions and normality of errors would be critical for small data sets; with more data, less is needed. Conditioning on the $\{X_i\}$ is a popular option.

The critical assumption is: if you intervene to set the value x for X on subject i in the first 'experiment', the Y-value responds according to (1) above: the disturbance $\delta_i(\omega)$ is unaffected by intervention. If you set x and y as the values for X and Y on subject i in the second experiment, the Z-value responds according to (2) above; again, $\varepsilon_i(\omega)$ is unaffected. In particular, the assignment by Nature of subjects to levels of X does not affect the δ's or ε's. Given this structure, the parameters a, b, c, d, e in (1)–(3) above can be estimated from nonexperimental data and used to predict the results of interventions: for instance, setting X to x and Y to y should make Z around $\hat{c} + \hat{d}x + \hat{e}y$.

Pearl says in § 6 that he gives a 'clear empirical interpretation' and 'operational meaning' to causal assumptions, and clarifies their 'empirical basis'. There are two ways to read this:

(i) assumptions that justify causal inference from regression have been stated quite sharply;

(ii) feasible methods have been provided for validating these assumptions, at least in certain examples.

The first assertion seems right, indeed, that is one of the main contributions of the paper. The second reading, which is probably not the intended one, would be a considerable over-statement. Invariance of errors under hypothetical interventions is a tall order. How can we test that $Z_i(\omega)$ would have been $g(x, y) + \varepsilon_i(\omega)$ if only $X_i(\omega)$ had been set to x and $Y_i(\omega)$ to y? What about the stochastic assumptions on δ and ε? In the typical observational study, there is no manipulation of variables and precious little sampling. Validation of causal models remains an unresolved problem.

Pearl's framework is more general than equations (1)–(3) above, and the results are more subtle. Still, the causal laws, i.e. the analogues of the equations, are assumed rather than inferred from the data. One technical complication should be noted: in Pearl's equation (3), distributions are identifiable but the 'link functions' f_i are not. The focus is qualitative rather than quantitative, so weaker invariance assumptions may suffice. More discussion of this point would be welcome.

Concomitants. Concomitant variables pose further difficulties (Dawid, 1979). Thus, in equations (1)–(3) above, suppose X is a dummy variable for sex, Y is education and Z is income. Some would consider a counter-factual interpretation: How much would Harriet have made if she had been Harry and gone to college? Others would view X as not manipulable, even in principle. Setting a subject's sex to M, even in a thought experiment, is then beside the point. Robins (1986, 1987a) offers another way to model concomitants.

Conclusions. Pearl has developed mathematical language in which causal assumptions can be discussed. The gain in clarity is appreciable. The next step must be validation: to make real progress, those assumptions have to be tested.

[*Received April* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY GUIDO W. IMBENS

Department of Economics, Harvard University, Littauer Center, Cambridge, Massachusetts 02138, U.S.A.

AND DONALD B. RUBIN

Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, Massachusetts 02138, U.S.A.

Judea Pearl presents a framework for deriving causal estimands using graphical models representing statements of conditional independence in models with independent and identically distributed random variables, and a 'set' notation with associated rules to reflect causal manipulations. This is an innovative contribution as it formalises the use of path-analysis diagrams for causal inference, traditionally popular in many fields including econometrics, e.g. Goldberger (1973). Because Pearl's technically precise framework separates issues of identification and functional form, often inextricably linked in the structural equations of literature, this paper should serve to make this extensive literature more accessible to statisticians and reduce existing confusion between statisticians and econometricians: see, e.g., the Discussion of Wermuth (1992).

Our discussion, however, focuses on this framework as an alternative to the practice in statistics, typically based on the potential outcomes framework for causal effects, or Rubin causal model (Holland, 1986; Rubin, 1974, 1978), which itself is an extension of Neyman's (1923) formulation for randomised experiments as discussed by Rubin (1990). The success of these frameworks in defining causal estimands should be measured by their applicability and ease of formulating and assessing critical assumptions.

Much important subject-matter information is not conveniently represented by conditional independence in models with independent and identically distributed random variables. Suppose that, when a person's health status is 'good', there is no effect of a treatment on a final health outcome, but, when a person's health status is 'sick', there is an effect of this treatment, so there is dependence of final health status on treatment received conditional on initial health status. Although the available information is clearly relevant for the analysis, its incorporation, although immediate using potential outcomes, is not straightforward using graphical models.

Next, consider a two-treatment randomised experiment with imperfect compliance, so that the received treatment is, in general, not ignorable despite the ignorability of the assigned treatment. Assuming that any effect of the assigned treatment on the outcome works through the received treatment, one has the instrumental variables example in Fig. 5(b), which excludes a direct effect of Z on Y, given X. In other work ('Bayesian inference for causal effects in randomized experiments with noncompliance', Working Paper 1976, Harvard Institute of Economic Research, Harvard University) we have discussed important distinctions between different versions of this exclusion restriction, which can be stated using potential outcomes but are blurred in graphical models. In that paper and related work (Imbens & Angrist, 1994; Angrist, Imbens & Rubin, 1995), we also stress the importance of the 'monotonicity assumption', requiring the absence of units taking the treatment if assigned to control and not taking it if assigned to treatment. This allows identification of the average effect of the treatment for the subpopulation of compliers without assuming a common, additive, treatment effect for all units. Yet the monotonicity assumption is difficult to represent in a graphical model without expanding it beyond the representation in Fig. 5(b).

Complications also arise in Pearl's framework when attempting to represent standard experimental designs (Cochran & Cox, 1957) having clustering of units in nests, split-plot randomisations, carryover treatments, etc.

A related reason for preferring the Rubin causal model is its explicit distinction between assign-

ment mechanisms, which are often to some extent under the investigator's control even in observational studies, and scientific models underlying the data, which are not. Consider the discussion of the equivalence of the Rosenbaum–Rubin condition of strong ignorability of the assignment mechanism and the back-door criterion. In general, the concept of ignorable assignment (Rubin, 1976, 1978) does not require the conditional independence used in Pearl's analysis. For example, a sequential assignment mechanism with future treatment assignments dependent on observed outcomes of previous units is ignorable, but such an assignment mechanism apparently requires a very large graphical model with all units defined as separate nodes, thereby making Pearl's results, which require 'an arbitrary large sample randomly drawn from the joint distribution', irrelevant.

Finally, consider the smoking–tar–cancer example in Figs 3, 4 and 6(e). Pearl claims that his analysis reveals that one can learn the effect of one's smoking on one's lung cancer from observational data, the only provision being that 'smoking does not have any direct effect on lung cancer except that mediated by tar deposits', i.e. no direct arrow from X to Y. But this claim is misleading as there are many other provisions hidden by the lack of an arrow between X and Z. For example, suppose that smokers are more likely to live in cities, and therefore more likely to be exposed to tar through pollution, or that smokers are more likely to interact with smokers, and are therefore exposed to more second-hand smoke than nonsmokers, etc. In this example the role of 'tar deposits' as an outcome is confounded with its role as a cause whose assignment may partially depend on previous treatments and outcomes, as can occur in serial experiments (Herzberg & Cox, 1969).

Our overall view of Pearl's framework is summarised by Hill's concluding sentence (1971, p. 296). 'Technical skills, like fire, can be an admirable servant and a dangerous master'. We feel that Pearl's methods, although formidable tools for manipulating directed acyclical graphs, can easily lull the researcher into a false sense of confidence in the resulting causal conclusions. Consequently, until we see convincing applications of Pearl's approach to substantive questions, we remain somewhat sceptical about its general applicability as a conceptual framework for causal inference in practice.

[*Received April* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY JAMES M. ROBINS

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

1. INTRODUCTION

Pearl has carried out two tasks. In the first, in § 2, using some results of Spirtes, Glymour & Scheines (1993), he showed that a nonparametric structural equations model depicted as a directed acyclic graph G implies that the causal effect of any variables $X \subseteq G$ on $Y \subseteq G$ is a functional of (i) the distribution function P_G of the variables in G, and (ii) the partial ordering of these variables induced by the directed graph. This functional is the g-computation algorithm functional, hereafter g-functional, of Robins (1986, p. 1423). In the second, in §§ 3–5, only a subset of the variables in G is observed. Given known conditional independence restrictions on P_G encoded as missing arrows on G, Pearl develops elegant graphical inference rules for determining identifiability of the g-functional from the law of the observed subset. Task 2 requires no reference to structural models or to causality. A potential problem with Pearl's formulation is that his structural model implies that all variables in G, including concomitants such as age or sex, are potentially manipulable. Below I describe a less restrictive model that avoids this problem but, when true, still implies that the g-functional equals the effect of the treatments X of interest on Y. This critique of Pearl's structural model is unconnected with his graphical inference rules, which were his main focus and

are remarkable and path-breaking, going far beyond my own and others' results (Robins, 1986, § 8, Appendix F).

2. TASK 1

2·1. *General*

Robins (1986, 1987b) proposed a set of counterfactual causal models based on event trees, called causally interpreted structured tree graphs, hereafter causal graphs, that includes Pearl's nonparametric structural equations model as a special case. These models extended Rubin's (1978) 'time-independent treatment' model to studies with direct and indirect effects and time-varying treatments, concomitants, and outcomes. In this section, I describe some of these models.

2·2. *A causal model*

Let $V_i = \{V_{1i}, \ldots, V_{Mi}\}$ denote a set of temporally-ordered discrete random variables observed on the ith study subject, $i = 1, \ldots, n$. Let $X_i := \{X_{1i}, \ldots, X_{Ki}\} \subseteq V_i$ be temporally-ordered, potentially manipulable, treatment variables of interest. The effect of X_i on outcomes $Y_i \subseteq V_i \backslash X_i$ is defined to be $\text{pr}\{Y_i(x) = y\}$, where the counterfactual random variable $Y_i(x)$ denotes a subject's Y value had all n subjects followed the generalised treatment regime $g = x := \{x_1, \ldots, x_K\}$. Robins (1986) wrote $\text{pr}\{Y_i(x) = y\}$ as $\text{pr}(y|g = x)$. Pearl substitutes $\text{pr}(y|\check{x})$. We regard the

$$\{V_i,\ Y_i(x);\ x \in \text{support of } X_i\} \quad (i = 1, \ldots, n)$$

as independent and identically distributed, and henceforth suppress the i subscript.

This formal set-up can accommodate a superpopulation model with deterministic outcomes and counterfactuals as did that of Rubin (1978). Suppose we regard the n study subjects as randomly sampled without replacement from a large superpopulation of N subjects, and our interest is in the causal effect of X on Y in the superpopulation. Then, even if for each superpopulation member, V and $Y(x)$ were deterministic nonrandom quantities, nonetheless, in the limit as $N \to \infty$ and $n/N \to 0$, we can model the data on the n study subjects as independent and identically distributed draws from the empirical distribution of the superpopulation.

We now show that $\text{pr}(y|g = x)$ is identified from the law of V if each component X_k of X is assigned at random given the past. Let L_k be the variables occurring between X_{k-1} and X_k, with L_1 being the variables preceding X_1. Write $\bar{L}_k := (L_1, \ldots, L_k)$, $L := \bar{L}_K$ and $\bar{X}_k := (X_1, \ldots, X_k)$, and define $\bar{X}_0$, $\bar{L}_0$ and $\bar{V}_0$ to be identically 0. In considering Task 1 I have proved the following (Robins, 1987b, Theorem AD.1 and its corollary).

THEOREM. *If, in Dawid's* (1979) *conditional independence notation, for all k,*

$$Y(x) \perp\!\!\!\perp X_k | \bar{L}_k, \bar{X}_{k-1} = \bar{x}_{k-1}, \tag{1}$$

$$X = x \Rightarrow Y(x) = Y, \tag{2}$$

$$\text{pr}(X_k = x_k | \bar{X}_{k-1} = \bar{x}_{k-1}, \bar{L}_k) \neq 0, \tag{3}$$

then

$$\text{pr}(y|g = x) = h(y|g = x), \tag{4}$$

where

$$h(y|g = x) := \sum_{\bar{l}_K} \text{pr}(y|\bar{l}_K, \bar{x}_K) \prod_{k=1}^{K} \text{pr}(l_k|\bar{l}_{k-1}, \bar{x}_{k-1})$$

is the g-functional for x on y based on covariates L. If X is univariate,

$$h(y|g = x) = \sum_{l_1} \text{pr}(y|x, l_1)\,\text{pr}(l_1)$$

(*Rosenbaum & Rubin*, 1983).

Following Robins (1987b, p. 327), I shall refer to V as a $R(Y, g = x)$ causal graph whenever (1) and (2) above hold, where $R(Y, g = x)$ stands for 'randomised with respect to Y for treatment $g = x$ given covariates L'. Robins et al. (1992) called (1) the assumption of no unmeasured confounders given L. Under the aforementioned superpopulation model, (1) will hold in a true sequential randomised trial with X randomised and X_k-specific randomisation probabilities that depend only on the past $(\bar{L}_k, \bar{X}_{k-1})$. In observational studies, (1) is untestable; investigators can at best hope to identify covariates L so that (1) is approximately true. Equation (2) is Rubin's (1978) stable unit treatment value assumption: it says Y and $Y(x)$ are equal for subjects with $X = x$, irrespective of other subjects' X values. Robins (1993) shows that

$$h(y|g = x) = E\left\{I(X = x)I(Y = y) \Big/ \prod_{k=1}^{K} \mathrm{pr}(x_k|\bar{x}_{k-1}, \bar{L}_k)\right\},$$

whose denominator clarifies the need for (3). See also Rosenbaum & Rubin (1983).

2·3. *Relationship with Pearl's work*

Suppose we represent our ordered variables $V = \{V_1, \ldots, V_M\}$ by a directed acyclic graph G that has no missing arrows, so that $\bar{V}_{m-1} := (V_1, \ldots, V_{m-1})$ are V_m's parents. Then Pearl's nonparametric structural equation model becomes

$$V_m = f_m(\bar{V}_{m-1}, \varepsilon_m), \tag{5}$$

for $f_m(.,.)$ unrestricted $(m = 1, \ldots, M)$, and

$$\varepsilon_m \quad (1 \leqslant m \leqslant M) \tag{6}$$

are jointly independent.

Pearl's assumption of missing arrows on G is (i) more restrictive than (5), and (ii) only relevant when faced with unobserved variables, as in Task 2. We now establish the equivalence between model (5)–(6) above and a particular causal graph, the finest fully randomised causal graph. For any $X \subset V$, $x \in$ support X, let the counterfactual random variable $V_m(x)$ denote the value of V_m had X been manipulated to x.

DEFINITIONS (*Robins*, 1986, *pp.* 1421–2). (*a*) *We have that V is a finest causal graph if (i) all one-step ahead counterfactuals $V_m(\bar{v}_{m-1})$ exist, and (ii) V and the counterfactuals $V_m(x)$ for any $X \subset V$ are obtained by recursive substitution from the $V_m(\bar{v}_{m-1})$; for example*

$$V_3 \equiv V_3\{V_1, V_2(V_1)\}, \quad V_3(v_1) = V_3\{v_1, V_2(v_1)\}.$$

(*b*) *A finest causal graph V is a finest fully randomised causal graph if, for all m,*

$$\{V_{m+1}(\bar{V}_{m-1}, v_m), \ldots, V_M(\bar{V}_{m-1}, v_m, \ldots, v_{M-1})\} \perp\!\!\!\perp V_m | \bar{V}_{m-1}. \tag{7}$$

For V to be a finest causal graph, all variables $V_m \in V$ must be manipulable. Equation (7) above essentially says that each V_m was assigned at random given the past $\bar{V}_{m-1}$. In particular, (7) would hold in a sequential randomised trial in which all variables in V, not just the treatments X of interest, are randomly assigned given the past.

LEMMA 1. (i) *Equation* (5) *above is equivalent to V's being a finest causal graph, and* (*ii*) *equations* (5) *and* (6) *above are jointly equivalent to V's being a finest fully randomised causal graph.*

Proof of Lemma. If (5) holds, define $V_m(\bar{v}_{m-1})$ to be $f_m(\bar{v}_{m-1}, \varepsilon_m)$. Conversely, given $V_m(\bar{v}_{m-1})$, define $\varepsilon_m = \{V_m(\bar{v}_{m-1})$: $\bar{v}_{m-1} \in$ support of $\bar{V}_{m-1}\}$ and set $f_m(\bar{v}_{m-1}, \varepsilon_m) = V_m(\bar{v}_{m-1})$. Part (ii) follows by some probability calculations. □

The statement 'V a finest fully randomised causal graph' implies that V is a $R(Y, g = x)$ causal graph, and thus, given (3) above, that $\mathrm{pr}(y|g = x) = h(y|g = x)$. The converse is false. For example, 'V a $R(Y, g = x)$ causal graph' only requires that the treatments X of interest be manipulable.

3. Task 2

Given (1)–(3) above, to obtain $\mathrm{pr}(y|g=x)$, we must compute $h(y|g=x)$. However, often data cannot be collected on a subset of the covariates $L \subseteq V$ believed sufficient to make (1) above approximately true. Given a set of correct conditional independence restrictions on the law of V, encoded as missing arrows on a directed acyclic graph G over V, Pearl provides graphical inference rules for determining whether $h(y|g=x)$ is identified from the observed data. Pearl's graphical inference rules are correct without reference to counterfactuals or causality when we define $\mathrm{pr}(y|\check{x}, \check{z}, w)$ to be

$$h\{y, w|g=(x, z)\}/h\{w|g=(x, z)\}.$$

Unfortunately, since covariates are missing, an investigator must rely on often shaky subject matter beliefs to guide link-deletions. Pearl & Verma (1991) appear to argue, although I would not fully agree, that beliefs about causal associations are generally sharper and more accurate than those about noncausal associations. If so, it would be advantageous to have all potential links on G represent direct causal effects, which will be the case only if V is a finest fully randomised causal graph and would justify Pearl's focus on nonparametric structural equation models.

[*Received April* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

By PAUL R. ROSENBAUM

Department of Statistics, University of Pennsylvania, 3000 Steinberg Hall-Dietrich Hall, Philadelphia, Pennsylvania 19104-6302, U.S.A.

1. Successful and unsuccessful causal inference: Some examples

Example 1. Cameron & Pauling (1976) gave vitamin C to patients with advanced cancers and compared survival to untreated controls. They wrote: 'Even though no formal process of randomisation was carried out ... we believe that [treated and control groups] come close to representing random subpopulations', expressing their belief in the following diagram.

(Treatment) → (Survival)

They concluded: '... there is strong evidence that treatment ... [with vitamin C] ... increases survival time'. Moertel et al. (1985) repeated this in a randomised trial, but found no evidence that vitamin C prolongs survival. Today, few believe vitamin C is effective against cancer. The studies have the same path diagram, but only the randomised trial gave the correct inference.

Example 2. The Coronary Drug Project compared lipid-lowering drugs, including clofibrate, to placebo in a randomised trial (May et al., 1981). We focus on the comparison of placebo and clofibrate. A drug can work only if consumed, yielding the following diagram.

(Assigned clofibrate or placebo) → (Amount of clofibrate consumed) → (Survival)

In the clofibrate group, the Project found 15% mortality at five years among good compliers who took their assigned clofibrate as opposed to 25% mortality among poor compliers who did not take their assigned clofibrate. Theorem 2 suggests clofibrate prolongs survival. Alas, it does not. In the placebo group, the mortality rates among good compliers who took their placebo was 15% compared to 28% mortality among poor compliers who did not take their placebo. Total mortality was similar in the entire clofibrate and placebo groups. Again, the nonrandomised comparison of level of clofibrate gave the wrong inference while the randomised comparison of entire clofibrate and placebo groups gave the correct inference.

Definition 2 is not a definition of causal effect, but rather an enormous web of assumptions. It asserts that a certain mathematical operation, namely this wiping out of equations and fixing of variables, predicts a certain physical reality, namely how changes in treatments, programmes and policies will change outcomes. No basis is given for believing that physical reality behaves this way. The examples above suggest it does not. See also Box (1966).

2. Warranted inferences

We do not say an inference is justified because it depends upon assumptions. We distinguish warranted and unwarranted inferences. To say, as Fisher (1935) said, that randomisation is the 'reasoned basis for inference' is to say it warrants a particular causal inference; a warrant is a reasoned basis. An assumption is not a basis for inference unless the assumption is warranted. Path diagrams allow one to make a large number of complex, interconnected assumptions, but this is not desirable, because it is much more difficult to ensure that the assumptions are warranted.

Inferences about treatment effects can sometimes be warranted by the following methods.

(i) Care in research design, for instance random assignment of treatments, may provide a warrant.

(ii) Insensitivity to substantial violations of assumptions may provide a warrant. For instance, the conclusion that heavy smoking causes lung cancer is highly insensitive to the assumption that smokers are comparable to nonsmokers (Cornfield et al., 1959; Rosenbaum, 1993, 1995).

(iii) Confirmation of numerous, elaborate predictions of a simple causal theory may at times provide a warrant. Here is Fisher's advice, as discussed by Cochran (1965, § 5):

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these is found to hold.

This advice is quite the opposite of finding the conditions that just barely identify a path model. Fisher is calling for a simple theory that makes extensive, elaborate predictions each of which can be contrasted with observable data to check the theory, that is an extremely overidentified model. See Rosenbaum (1984a, 1995) for related theory and practical examples.

[*Received May* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

By GLENN SHAFER

Faculty of Management, Rutgers University, Newark, New Jersey 07102, U.S.A.

This is an innovative and useful paper. It establishes a framework in which both probability and causality have a place, and it uses this framework to unify and extend methods of causal inference developed in several branches of statistics.

Pearl's framework is the graphical model. He brings probability and causality together by giving this model two roles: (i) it expresses a joint probability distribution for a set of variables, and (ii) it tells how interventions can change these probabilities. I find this informative and attractive. When it fits a problem, it provides a clear understanding of causality. But how often does it fit? People tend to become uncomfortable as soon as we look at almost any extensive example. Even in Pearl's own examples, it is hard to agree that each causal connection is equivalent to an opportunity for intervention, or that the simplest interventions are those that remove a particular variable from the

mechanism. If we try to do something about the birds, it will surely fall short of fixing their number at a desired level, and it may have other effects on the yield of the crop.

My inability to overcome objections of these kinds when I defend causal claims made for graphical models has led me to undertake a more fundamental analysis of causality in terms of probability trees. This analysis, which will be reported in a forthcoming book (Shafer, 1996), opens the way to generalising Pearl's ideas beyond their over-reliance on the idea of intervention.

A probability tree is causal if it is nature's tree i.e. if it shows how things happen step by step in nature. Causes are represented by steps in the tree. These steps determine the overall outcome, i.e. the path nature takes through the tree, and hence every variable. Some steps identify opportunities for intervention, but others simply represent how the world works. Variables are not causes, but they can be causally related. For example, two variables are independent in the probability-tree sense if they have no common causes: there is no step in the tree where both their probability distributions change. This implies that the variables are independent in the usual sense at every point in the tree. Similarly, two numerical variables are uncorrelated in the probability-tree sense if there is no step where both their expected values change, and this implies that they are uncorrelated in the usual sense at every point in the tree.

Pearl's graphical-model assumptions, explicit and implicit, correspond to the following statements about nature's probability tree: (i) if $i < j$, then X_i is settled before X_j, and (ii) at any point in the tree where X_{i-1} is just settled, the probability of X_i eventually coming out equal to x_i is $p(x_i | pa_i)$, where pa_i is the value of X_i's parents. These two conditions imply that Pearl's conditional independence relations, that each variable is independent of its nondescendants given its parents, hold at every point in the tree.

What is Pearl's $p(y | \check{x}_i)$ in probability-tree terms? It is an average of probabilities: we look at each point in the tree where X_{i-1} has just been settled, find the point following where X_i is settled to have the value x_i, and find the probability at that point that Y will come out equal to y. Then we average these probabilities of y, weighting each by the probability of the point where X_{i-1} was settled. This average tells us something about how steps in the direction of x_i, after X_{i-1} is settled, tend to promote y. It has causal meaning, for it describes how the world works, but it does not depend how well steps between X_{i-1} and x_i can be targeted by human intervention.

This idea of using averages to summarise the causal effect of steps in a probability tree does not depend on Pearl's graphical-model assumptions. What is needed, in general, is some way of describing a cut across nature's tree, in addition to the event or variable that identifies the following steps whose effect we want to average. In observational studies, the cut is often specified by choosing concomitants that are just settled there. In randomised studies, it can be specified by the conditions of the experiment, without explicit measurement.

Pearl's ideas can also be used in graphical models that make weaker assumptions about nature's tree. In particular, they can be used in path models, which represent only changes in expected values, not all changes in probabilities. These models are rather more flexible than Pearl's graphical models, contrary to the suggestion conveyed by Pearl's use of the term 'nonparametric'.

[*Received April* 1995]

Discussion of 'Causal diagrams for empirical research' by J. Pearl

BY MICHAEL E. SOBEL

Department of Sociology, University of Arizona, Tucson, Arizona 85721, U.S.A.

1. INTRODUCTION

Pearl takes the view, widely held in the social and behavioural sciences, that structural equation models are useful for estimating effects that correspond to those obtained if a randomised or

conditionally randomised experiment were conducted. Typically, linear models are used, and parameters or functions of these interpreted as unit or average effects, direct or total. A few workers argue if endogenous variables are viewed as causes, these should be treated as exogenous, and a new hypothetical system, absent equations for these variables, considered. As the effects are defined in the new system, assumptions about the relationship between parameters of the old pre-intervention and new post-intervention systems are needed (Sobel, 1990).

Pearl neatly extends this argument. His equation (5) specifies the post-intervention probabilities $\mathrm{pr}(x_1, \ldots, x_n | \check{x}_i')$ in terms of the model based pre-intervention probabilities. The effects of X_i on X_j are comparisons of $\mathrm{pr}(x_j | \check{x}_i')$ with $\mathrm{pr}(x_j | \check{x}_i^*)$, where x_i' and x_i^* are distinct values of X_i. If $(X_1, \ldots, X_n)$ is observed, $\mathrm{pr}(x_j | \check{x}_i')$ is identified; Pearl considers the nontrivial case, giving sufficient conditions for identifiability.

Pearl suggests his results are equivalent to those in Rubin's model for causal inference. For example, he claims that the back-door criterion is 'equivalent to the ignorability condition of Rosenbaum & Rubin (1983)'; if so, it should follow that

$$\mathrm{pr}(x_j | \check{x}_i', w) = \mathrm{pr}(x_{j_{x_i'}} | w),$$

the probability if all units in the subpopulation $W = w$ take value x_i' of the cause. Thus,

$$\mathrm{pr}(x_j | \check{x}_i') = \mathrm{pr}(x_{j_{x_i'}}).$$

But strong ignorability, given covariates W, implies

$$\mathrm{pr}(x_j | x_i', w) = \mathrm{pr}(x_{j_{x_i'}} | w);$$

the back-door criterion implies $\mathrm{pr}(x_j | x_i', w) = \mathrm{pr}(x_j | \check{x}_i', w)$. Neither condition implies the other; the two are only equivalent if

$$\mathrm{pr}(x_{j_{x_i'}} | w) = \mathrm{pr}(x_j | \check{x}_i', w).$$

The assumption

$$\mathrm{pr}(x_{j_{x_i'}} | w) = \mathrm{pr}(x_j | \check{x}_i', w),$$

and others like these, is the basis on which the paper rests, and for the equivalence claims made; such quantities need not be identical.

Suppose ignorability and the back-door criterion hold above. Then

$$\mathrm{pr}(x_{j_{x_i'}} | w) = \mathrm{pr}(x_j | \check{x}_i', w);$$

equality is now a conclusion. If W is observed, ignorability implies

$$\mathrm{pr}(x_{j_{x_i'}} | w) = \mathrm{pr}(x_j | x_i', w),$$

which can be computed directly. The problematic case in observational studies occurs when some covariates are unobserved. But supplementing ignorability with assumptions like those in this paper helps to identify effects in Rubin's model in such cases. To simplify, only atomic interventions will be considered and the outcome treated as a scalar quantity. Only the back-door criterion is considered, but Theorem 2, for example, could also be handled.

2. Ignorability and the back-door criterion

For an atomic intervention, rule 2, which supports the back-door criterion, is

$$\mathrm{pr}(x_j | \check{x}_i', w) = \mathrm{pr}(x_j | x_i', w) \tag{1}$$

if $(X_j \perp\!\!\!\perp X_i | W)_{G_{\underline{X}_i}}$. Assume, following Pearl's discussion of the back-door criterion, that W is observed. Ostensibly, (1) looks similar to the assumption of strongly ignorable treatment assignment: $X_{j_{x_i}} \perp\!\!\!\perp X_i | W$ for all values x_i of X_i, $0 < \mathrm{pr}(X_i = x_i | W = w)$ for all (x_i, w).

LEMMA 1. *Equality in* (1) *holds if* $PA_i \not\subseteq W$,

$$X_j \perp\!\!\!\perp (PA_i \backslash W) | (X_i, W). \tag{2}$$

Proof. This follows from

$$\text{pr}(x_j, w | \breve{x}_i') = \sum_{pa_i \backslash w} \frac{\text{pr}(x_j | w, x_i', pa_i \backslash w)\, \text{pr}(w, x_i', pa_i \backslash w)}{\text{pr}(x_i' | pa_i)}. \tag{3}$$

□

LEMMA 2. *If* $PA_i \not\subseteq W$, *the independence conditions in* (1) *and* (2) *are equivalent.*

Proof. Suppose (2) holds in G. In $G_{\underline{X}_i}$, any path p from X_i to X_j has the form $X_i \leftarrow M \ldots \rightarrow X_j$, where $M \in PA_i$. If $M \in PA_i \backslash W$, the subpath $M \ldots \rightarrow X_j$ in G is blocked by W, hence p is blocked by W in $G_{\underline{X}_i}$. Otherwise, p is a subpath of $l \rightarrow X_i \leftarrow M \ldots \rightarrow X_j$ in G, $l \in PA_i \backslash W$, which is blocked by W as it has arrows converging to X_i, implying p is blocked by W in $G_{\underline{X}_i}$. Conversely, in G any path p^* from l to X_j has form (a) $l \ldots \rightarrow X_i \rightarrow \ldots \rightarrow X_j$, (b) $l \ldots \rightarrow X_i \leftarrow M \ldots \rightarrow X_j$, or (c) $l \ldots \rightarrow X_j$, where X_i does not appear. Here X_i blocks type (a) paths. Type (b) paths contain subpaths $X_i \leftarrow M \ldots \rightarrow X_j$; by hypothesis, W blocks these in $G_{\underline{X}_i}$, implying W blocks p^* in G. Type (c) paths are subpaths of $X_i \leftarrow l \ldots \rightarrow X_j$; by hypothesis, W blocks these in $G_{\underline{X}_i}$, implying W blocks p^* in G. □

THEOREM. *If treatment assignment is strongly ignorable, given* $(W, PA_i \backslash W)$, *where* $PA_i \backslash W \neq \varnothing$ *is unobserved,* (2) *above holds, and no node in* W *is a descendant of* X_i, $\text{pr}(x_{j_{x_i}})$ *is identified, and* $\text{pr}(x_{j_{x_i}}) = \text{pr}(x_j | \breve{x}_i')$.

Proof. We have

$$\text{pr}(x_j | x_i', w, pa_i \backslash w) = \text{pr}(x_{j_{x_i}} | w, pa_i \backslash w), \tag{4}$$

$$\text{pr}(x_j | x_i', w, pa_i \backslash w) = \text{pr}(x_j | x_i', w), \tag{5}$$

where (4) follows from strong ignorability and (5) from (2); jointly, (4) and (5) imply

$$\text{pr}(x_{j_{x_i}} | w, pa_i \backslash w) = \text{pr}(x_{j_{x_i}} | w) = \text{pr}(x_j | x_i', w). \tag{6}$$

Since no node in W is a descendant of X_i,

$$\text{pr}(x_{j_{x_i}}) = \sum_w \text{pr}(x_j | x_i', w)\, \text{pr}(w) = \text{pr}(x_j | \breve{x}_i'). \qquad \square$$

[*Received May* 1995]

Rejoinder to Discussions of 'Causal diagrams for empirical research'

BY JUDEA PEARL

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

1. GENERAL

The subject of causality seems inevitably to provoke controversy among scientists, perhaps because causality is so essential to our thought and yet so seldom discussed in the technical

literature. I am pleased, therefore, to have the opportunity to respond to a range of concerns about the usefulness of the ideas developed in my paper.

2. Graphs, structural equations and counterfactuals

Underlying many of the discussions are queries about the assumptions, power and limitations of the three major notational schemes used in causal analysis: structural equations, graphs and the Neyman–Rubin–Holland model, henceforth called 'counterfactual analysis'. Thus, it seems useful to begin by explicating the commonalities among the three representational schemes, as noted in the Discussions of Freedman, following Holland (1988), Robins and Sobel; I will start with a structural interpretation of counterfactual sentences and then provide a general translation from graphs back to counterfactuals.

The primitive object of analysis in the 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 a natural interpretation in structural equations model. Consider a set T of equations

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

where the U_i are latent exogenous variables or disturbances and the PA_i are the observed explanatory variables. Equation (1) above is similar to (3) in my paper, except we no longer insist on the equations being recursive or on the U_i's being independent. Let U stand for the vectors $(U_1, \ldots, U_n)$, 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

$$Y(x, u) := Y_{T_x}(u), \tag{2}$$

namely, $Y(x, u)$ is the unique solution for Y under the realisation $U = u$ 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 characterise 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 modelling. Equation (2) above 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 by $Y(x)$ or Y_x. The counterfactual analysis proceeds by imagining the observed distribution $\mathrm{pr}(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function pr* defined over both observed and counterfactual variables. Queries about causal effects, written $\mathrm{pr}(y|\check{x})$ in the structural analysis, are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $\mathrm{pr}\{Y(x) = y\}$. The new entities $Y(x)$ are treated as ordinary random variables that are connected to the observed variables via the logical constraints (Robins, 1987b)

$$X = x \Rightarrow Y(x) = Y \tag{3}$$

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

For example, to communicate the understanding that in a randomised clinical trial, see Fig. 5(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 process is randomised, hence independent of any variation in the treatment selection process, structurally written as $U_X \perp\!\!\!\perp U_Z$, 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, $\text{pr}\{Y(x) = y\}$; 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 systematised. 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 systematised using the following translation from graphs.

Graphs provide qualitative information about the structure of both the equations in the model and the probability function $\text{pr}(u)$. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model (1) above. Additionally, 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 and $\{U_{Z_1}, \ldots, U_{Z_k}\}$, are independent in $\text{pr}(u)$. 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.

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

$$Y(pa_Y) = Y(pa_Y, s). \tag{4}$$

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

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\}. \tag{5}$$

For example, the graph in Fig. 3, displaying the parent sets

$$PA_X = \{\varnothing\}, \quad PA_Z = \{X\}, \quad PA_Y = \{Z\}$$

encodes the following assumptions.

Assumption 1: *Exclusion restrictions.* We require

$$X(z) = X(y) = X(z, y) = X(\varnothing) := X,$$

$$Z(y, x) = Z(x), \quad Y(z) = Y(z, x).$$

Assumption 2: *Independence restrictions.* We require

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

While it is not easy to see that these assumptions suffice for computing the causal effect $\text{pr}\{Y(x) = y\}$ using standard probability calculus together with axiom (3) above, the identifiability of $\text{pr}(y|\check{x})$ in the diagram of Fig. 3 ensures this sufficiency.

In summary, the structural and counterfactual frameworks are complementary to each other. Structural analysts can interpret counterfactual sentences as constraints over the solution set of a given system of equations (2) above and, conversely, counterfactual analysts can use the constraints over pr* given by (4) and (5) above as a definition of the graphs, structural equations and the physical processes which they represent.

3. The equivalence of counterfactual and structural analyses

Robins' discussion provides a concrete demonstration of the equivalence of the counterfactual and structural definitions of causal effects, $\text{pr}\{Y(x) = y\}$ and $\text{pr}(y|\check{x})$, respectively. Whereas (2)

above explicates counterfactual sentences in terms of operations on structural equations, Robins has done the converse by explicating the assumptions of a certain structural equations model in terms of counterfactual specifications. Specifically, starting with a complete directed acyclic graph with no confounding arcs, Robins translates the assumptions of error-independence in my (3) into the ignorability-type assumptions of his (1), and shows that causal effects can be expressed in the form of the *g*-functional in his (4), in full conformity with the post-intervention distribution in (5) in § 2·2. Note that in the structural equations framework the identifiability of causal effects in model (3) in my paper is almost definitional, because the post-intervention distribution (5) in § 2·2 follows immediately from the definition of an atomic intervention (Definition 2) and from the fact that deleting equations does not change the Markovian nature of (3), and hence the product form (2) applies. What is remarkable is that Robins has derived the same expression using counterfactual analysis, which, at least on the surface, is oblivious to meta-probabilistic notions such as equation deletion or error independence.

Robins' approach to dealing with missing links and unmeasured variables is different from mine; it follows the algebraic reduction method illustrated in § 3·2. After writing the *g*-functional using both observed and unobserved variables as in my (7) and (8), Robins would attempt to use the independencies embodied in $\mathrm{pr}(v)$ to eliminate the unobservables from the *g*-formula. Because the elimination only requires knowledge of the conditional independencies embodied in $\mathrm{pr}(v)$, any dependency-equivalent graph of G can be used or, for that matter, any nongraphical encoding of those independencies, for example, $\mathrm{pr}(v)$ itself. The price paid for this generality is complexity: many latent variables are being summed over unnecessarily, and I am not aware of any systematic way of eliminating the latent variables from the *g*-expression; see the transition from my (8) to (9).

The aim of §§ 4 and 5 of my paper is to demonstrate how the derivation can be systematised and simplified by abandoning this route and resorting instead to syntactic manipulation of formulae involving observed variables only. The derivation is guided by various subgraphs of G that depend critically on the causal directionality of the arrows, hence the conditional independencies carried by G will not suffice. It is quite possible that some of these manipulations could be translated to equivalent operations on probability distributions but, if we accept the paradigm that the bulk of scientific knowledge is organised in the form of qualitative causal models rather than probability distributions, I do not see tremendous benefit in such effort.

Sobel is correct in pointing out that the equivalence of the ignorability and the back-door conditions hinges upon the equality $\mathrm{pr}\{Y(x) = y\} = \mathrm{pr}(y|\dot{x})$. Robins' results and the translations of (4) and (5) above provide the basis for this equality. I am puzzled, though, by Rosenbaum's astonishment at the possibility that 'a certain mathematical operation, namely this wiping out of equations ..., predicts a certain physical reality'. While it may seem odd that post-Galilean scientists habitually expect reality to obey the predictions of mathematical operations, the perfect match between mathematical predictions based on Definition 2 and those obtained by other, less manageable approaches reaffirms the wisdom of this expectation; the scientific basis for deleting equations is given in the paragraph preceding Definition 2.

4. Practical versus hypothetical interventions

Freedman's concern that invariance of errors under interventions may be a 'tall order' is a valid concern when addressed to practical, not to hypothetical interventions. Given a structural equation $Y = f(X, U)$, the hypothetical atomic intervention $\mathrm{set}(X = x)$ always leaves f and U invariant, by definition; see (2) above. The crucial point is that, in order to draw valid inferences about the effect of physically fixing X at x, we must assume that our means of fixing X possesses the local property of the operator $\mathrm{set}(X = x)$, that is it affects only the mechanism controlling X, and leaves all other mechanisms, e.g. the function f, unaltered. If current technology is such that every known method of fixing X produces side effects, then those side effects should be specified and modelled as conjunctions of several atomic interventions. Naturally, causal theories can say nothing about interventions that might break down every mechanism in the system in a manner unknown to the

modeller. Causal theories are about a class of interventions that affect a select set of mechanisms in a prescribed way.

Note that this locality assumption is tacitly embodied in every counterfactual utterance as well as in the counterfactual variable $Y(x)$ used in Rubin's model. When we say 'this patient would have survived had he taken the treatment', we exclude from consideration the eventuality that the patient takes the treatment but shoots himself. It is only by virtue of this locality assumption that we can predict the effect of practical interventions, e.g. how a patient would react to the legislated treatment, from counterfactual inferences about behaviour in a given experimental study.

Freedman's difficulty with unmanipulable concomitants such as age and sex is of a slightly different nature because, here, it seems that we lack the mental capacity to imagine even hypothetical interventions that would change these variables. Remarkably, however, people do not consider common expressions such as 'If you were younger' or 'Died from old age' to be as outrageous as manipulating one's age might suggest. Why? The answer, I believe, lies in the structural equations model of (1) and (2) above. If age X is truly nonmanipulable, then the process determining X is considered exogenous to the system and X is modelled as a component of U, or a root node in the graph. As such, no manipulation is required for envisioning the event $X = x$; we can substitute $X = x$ in U without deleting any equations from the model and obtain $\text{pr}(y|\hat{x}) = \text{pr}(y|x)$ for all x and y. Additionally, in employment discrimination cases, the focus of concern is not the effect of sex on salaries but rather the effect of the employer's awareness of the plaintiff's sex on salary. The latter effect is manipulable, both in principle and in practice.

Shafer's uneasiness with the manipulative account of causation also stems from taking the notion of intervention, too literally, to mean human intervention. In the process of setting up the structural equations (1) above or their graphical abstraction the analyst is instructed to imagine hypothetical interventions as defined by the submodel T_x in Definition 2 and equation (2) above, regardless of their feasibility. Such thought experiments, for example slowing down the moon's velocity and observing the effect on the tides, are feasible to anyone who possesses a model of the processes that operate in a given domain.

The analysis in my paper invokes such hypothetical local manipulations, and I mean them to be as delicate and incisive as theory will permit; it does not insist on technologically feasible manipulations which, as Shafer and Freedman point out, might cause undesired side effects. Structural equations models, counterfactual sentences, and Shafer's probability trees all invoke the same type of hypothetical scenarios, but I find an added clarity in imagining the desired scenario as triggered by some controlled wilful act, rather than by some uncontrolled natural phenomenon, e.g. the moon hitting a comet, which might have its own, undesired side effects, e.g. the comet creating its own effects on the tides.

I agree with Shafer that not every causal thought identifies opportunities for human intervention, but I would argue strongly that every causal thought is predicated upon some notion of a 'change'. Therefore, a theory of how mechanisms are changed, assembled, replaced and broken down, be it by humans or by Nature, is essential for causal thinking.

5. Intervention as conditionalisation

I agree with Dawid that my earlier formulation (Pearl, 1993b), which incorporates explicit policy variables in the graph and treats intervention as conditionalisation on those variables, has several advantages over the functional representation emphasised here. Fienberg, Glymour & Spirtes articulate similar sentiments. Nonetheless, I am pursuing the functional representation, partly because it is a more natural framework for thinking about data-generating processes and partly because it facilitates the identification of 'causes of effects', especially in nonrecursive systems.

Balke & Pearl (1994), for example, show that sharp informative bounds on 'causes of effects' can sometimes be obtained without identifying the functions f_i or the variables ε_i. Additionally, if we can assume the functional form of the equations, though not their parameters, then the standard econometric conditions of parameter identification are sufficient for consistently inferring 'causes

of effects'. Balke & Pearl (1995) demonstrate how linear, nonrecursive structural models can be used to estimate the probability that 'event $X = x$ is the cause for effect E', by computing the counterfactual probability that, given effect E and observations O, 'E would not have been realised, had X not been x'.

6. Testing versus using assumptions

Freedman's concern that 'finding the mathematical consequences of assumptions matters, but connecting assumptions to reality matters too' has also been voiced by other discussants, most notably Dawid and Rosenbaum. Testing hypotheses against data is indeed the basis of scientific inquiry, and my paper makes no attempt to minimise the importance of such tests. However, scientific progress also demands that we not re-test or re-validate all assumptions in every study but, rather, that we facilitate the transference of knowledge from one study to another, so that the conclusions of one study may be imposed as assumptions in the next. For example, the careful empirical work of Moertel et al. (1985), which, according to Rosenbaum's discussion, refuted the hypothesis that vitamin C is effective against cancer, should not be wasted. Instead, their results should be imposed, e.g. as a missing causal link, in future studies involving vitamin C and cancer patients, so as to enable the derivation of new causal inferences. The transference of such knowledge requires a language in which the causal relationship 'vitamin C does not affect survival' receives symbolic representation. Such a language, to the best of my knowledge, so far has not become part of standard statistical practice. Moreover, a language for stating assumptions is not very helpful if it is not accompanied by the mathematical machinery for quickly drawing conclusions from those assumptions or reasoning backward and isolating assumptions that need be tested, justified, or reconsidered. Facilitating such reasoning comprises the main advantage of the graphical framework.

7. Causation versus dependence

Cox & Wermuth welcome the development of graphical models but seem reluctant to use graphs for expressing substantive causal knowledge. For example, they refer to causal diagrams as 'a system of dependencies that can be represented by a directed acyclic graph'. I must note that my results do not generally hold in such a system of dependencies; they hold only in systems that represent causal processes of which statistical dependencies are but a surface phenomenon. Specifically, the missing links in these systems are defined by asymmetric exclusion restrictions, as in (4) above, not by conditional independencies. The difficulties that Smith (1957) encounters in defining admissible concomitants indeed epitomise the long-standing need for precise notational distinction between causal influences and statistical dependencies.

Another type of problem created by lack of such a distinction is exemplified by Cox & Wermuth's 'difficulties emphasised by Haavelmo many years ago'. These 'difficulties' are, see Discussions following Wermuth (1992) and Cox & Wermuth (1993): (i) the term ax in the structural equation $y = ax + \varepsilon$ normally does not stand for the conditional expectation $E(Y|x)$, and (ii) variables are excluded from the equation for reasons other than conditional independence. Haavelmo (1943), who emphasises these features in the context of nonrecursive equations, is very explicit about defining structural equations in terms of hypothetical experiments and, hence, does not view the difference between ax and $E(y|x)$ as a 'difficulty' of interpretation but rather as an important feature of a well-interpreted model, albeit one which requires a more elaborate estimation technique than least squares. Cox & Wermuth's difficulty stems from the reality that certain concepts in science do require both a causal and a probabilistic vocabulary. The many researchers who embrace this richer vocabulary, e.g. Haavelmo, find no difficulty with the interpretation of structural equations. I therefore concur with Imbens & Rubin's observation that the advent of causal diagrams should promote a greater understanding between statisticians and these researchers.

8. Exemplifying modelling errors

Rosenbaum mistakingly perceives path analysis as a competitor to randomised experiments and, in attempting to prove the former inferior, he commits precisely those errors that most path analysts have learned to avoid. After reporting a randomised study (Moertel et al., 1985) that gave different results from those of a nonrandomised study (Cameron & Pauling, 1976), he concludes that 'the studies have the same path diagram, but only the randomised trial gave the correct inference'. However, the two studies have different path diagrams. The diagram corresponding to the randomised trial is given in Fig. 6(a), while the diagram corresponding to the nonrandomised trial is shown in Fig. 7(a); the former is identifiable, the latter is not. Such modelling errors do not make the diagrams the same and do not invalidate the method.

In Rosenbaum's second example, with which he attempts to refute Theorem 2, he again introduces an incorrect diagram. The example involves a clinical trial in which compliance was imperfect, and the diagram corresponding to such trials is shown in Fig. 5(b). Because a confounding back-door path exists between X and Y, the conditions of Theorem 2 are not satisfied, and the causal effect is not identifiable: see the discussion in the second paragraph of § 5, and a full analysis of noncompliance given by Pearl (1995). The chain diagram chosen by Rosenbaum implies a conditional independence relation that does not hold in the data reported. Thus, Rosenbaum's attempted refutation of Theorem 2 is based on a convenient, but incorrect, diagram.

9. The myth of dangerous graphs

Imbens & Rubin perceive two dangers in using the graphical framework: (i) graphs hide assumptions; and (ii) graphs lull researchers into a false sense of confidence.

(i) Like all abstractions, graphs make certain features explicit while keeping details implicit, to be filled in by other means if the need arises. When an independence relationship does not obtain graphical representation, the information can be filled in from the numerical probabilities, or structural equations, that annotate the links of the graph. However, a graph never fails to display a dependency if the graph modeller perceives one; see (2) of my paper. Therefore, a graph analyst is protected from reaching invalid, unintended conclusions.

Imbens & Rubin's discussion of my smoking–tar–cancer example in Figs 3, 4 and 6(e) illustrate this point. Contrary to their statement, the provision that tar deposits not be confounded with smoking is not hidden in the graphical representation. Rather, it stands out as vividly as can be, in the form of a missing dashed arc between X and Z. I apologise that my terse summary gave the impression that a missing link between X and Y is the 'only provision' required. From the six provisions shown in the graph, I have elected to recall this particular one, but the vividness of the graph, condition (ii) of Theorem 2, equation (13), and the entire analysis, see also (4) and (5) above, should convince Imbens and Rubin that such provisions have not been neglected. In fact, graphs provide a powerful deterrent against forgetting assumptions unmatched by any other formalism. Every pair of nodes in the graph waves its own warning flag in front of the modeller's eyes: 'Have you neglected an arrow or a dashed arc?' I consider these warnings to be a strength, not a weakness, of the graphical framework.

(ii) Imbens & Rubin's distrust of graphs would suggest, by analogy, that it is dangerous to teach differential calculus to physics students lest they become so enchanted by the convenience of the mathematics that they overlook the assumptions. Whilst we occasionally meet discontinuous functions that do not admit the machinery of ordinary differential calculus, this does not make the calculus useless or harmful. Additionally, I do not think over-confidence is currently holding back progress in statistical causality. On the contrary, I believe that repeated warnings against confidence are mainly responsible for the neglect of causal analysis in statistical research, and that such warnings have already done more harm to statistics than graphs could ever do.

Finally, I would like to suggest that people will be careful with their assumptions if given a language that makes those assumptions and their implications transparent; moreover, when assumptions are transparent, they are likely to be widely discussed. No matter how powerful, a

notational system that does not accommodate an explicit representation of familiar processes will only inhibit people from formulating and assessing assumptions. As a result, instead of being brought into the light, critical assumptions tend to remain implicit or informal, and important problems of causal inference go unexplored. Indeed, the theory of causal inference has so far had only minor impact on rank-and-file researchers, on the methods presented in statistics textbooks, and on public policy-making. I sincerely hope graphical methods can help change this situation, both by uncovering tangible new results and by transferring causal analysis from the academic to the laboratory.

[*Received June* 1995]

Additional References

Balke, A. & Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11*, Ed. P. Besnard and S. Hanks, pp. 11–8. San Francisco, CA: Morgan Kaufmann.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Box, G. E. P. (1966). The use and abuse of regression. *Technometrics* **8**, 625–9.

Cameron, E. & Pauling, L. (1976). Supplemental ascorbate in the supportive treatment of cancer: prolongation of survival times in terminal human cancer. *Proc. Nat. Acad. Sci.* (USA), **73**, 3685–9.

Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *J. R. Statist. Soc.* A **128**, 134–55.

Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

Dawid, A. P. (1984). Statistical theory. The prequential approach (with Discussion). *J. R. Statist. Soc.* A **147**, 278–92.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with Discussion). *J. R. Statist. Soc.* B **53**, 79–109.

Fairfield Smith, H. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. *Biometrics* **13**, 282–308.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Freedman, D. (1991). Statistical models and shoe leather (with Discussion). In *Sociological Methodology 1991*, Ed. P. Marsden, Ch. 10. Washington, D.C.: American Sociological Association.

Freedman, D. (1995). Some issues in the foundation of statistics (with Discussion). *Foundat. Sci.* **1**, 19–83.

Goldberger, A. S. (1973). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.

Herzberg, A. M. & Cox, D. R. (1969). Recent work on design of experiments: A bibliography and a review. *J. R. Statist. Soc.* A **132**, 29–67.

Hill, A. B. (1971). *A Short Textbook of Medical Statistics*, 10th ed. Place: Lippincott.

Holland, P. (1986). Statistical and causal inference. *J. Am. Statist. Assoc.* **81**, 945–70.

May, G., DeMets, D., Friedman, L., Furberg, C. & Passamani, E. (1981). The randomized clinical trial: Bias in analysis. *Circulation* **64**, 669–73.

Moertel, C., Fleming, T., Creagan, E., Rubin, J., O'Connell, M. & Ames, M. (1985). High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New Engl. J. Med.* **312**, 137–41.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on Principles, Section 9. Transl. (1990) in *Statist. Sci.* **5**, 465–80.

Robins, J. M. (1987a). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic Dis.* **40**, Suppl. 2, 139S–161S.

Robins, J. M. (1987b). Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect'. *Comp. Math. Applic.* **14**, 923–45.

Robins, J. M. (1993). Analytic methods for estimating HIV treatment and cofactor effects. In *Methodological Issues of AIDS Mental Health Research*, Ed. D. G. Ostrow and R. Kessler, pp. 213–90. New York: Plenum.

Rosenbaum, P. R. (1984a). From association to causation in observational studies. *J. Am. Statist. Assoc.* **79**, 41–8.

Rosenbaum, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Am. Statist. Assoc.* **88**, 1250–3.

Rosenbaum, P. R. (1995). *Observational Studies*. New York: Springer-Verlag.

Rubin, D. B. (1976). Inference and missing data (with Discussion). *Biometrika* **63**, 581–92.

SHAFER, G. (1996). *The Art of Causal Conjecture*. Cambridge, MA: MIT Press.
SMITH, H. F. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariates. *Biometrics* **13**, 282–308.
VOVK, V. G. (1993). A logic of probability, with application to the foundations of statistics (with Discussion). *J. R. Statist. Soc.* B **55**, 317–51.