Errata for J. Pearl, Causality: Models, Reasoning, and Inference (2nd edition).
Changes marked in red were implemented in corrected reprint 2013.
Changes marked in green are planned for the next printing (updated July 27, 2020).

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521895606

First published 2000
8th printing 2008
Second edition 2009
Reprinted 2010
Reprinted with corrections 2013
Printed in the United States of America

A catalog record for this publication is available from the British Library.

The Library of Congress has cataloged the first edition as follows:

Pearl, Judea
Causality : models, reasoning, and inference / Judea Pearl.
p. cm.
ISBN 0-521-77362-8 (hardback)
1. Causation. 2. Probabilities. I. Title.
BD541.P43 2000
122 – dc21 99-042108

ISBN 978-0-521-89560-6 Hardback

Readers who wish to be first introduced to the nonmathematical aspects of causation are advised to start with the Epilogue and then to sweep through the other historical/conceptual parts of the book: Sections 1.1.1, 3.3.3, 4.5.3, 5.1, 5.4.1, 6.1, 7.2, 7.4, 7.5, 8.3, 9.1, 9.3, and 10.1. More formally driven readers, who may be anxious to delve directly into the mathematical aspects and computational tools, are advised to start with Section 7.1 and then to proceed as follows for tool building: Section 1.2, Chapter 3, Sections 4.2–4.4, Sections 5.2–5.3, Sections 6.2–6.3, Section 7.3, and Chapters 8–10.

I owe a great debt to many people who assisted me with this work. First, I would like to thank the members of the Cognitive Systems Laboratory at UCLA, whose work and ideas formed the basis of many of these sections: Alex Balke, Blai Bonet, David Chickering, Adnan Darwiche, Rina Dechter, David Galles, Hector Geffner, Dan Geiger, Moisés Goldszmidt, Jin Kim, Jin Tian, and Thomas Verma. Tom and Dan have proven some of the most basic theorems in causal graphs; Hector, Adnan, and Moisés were responsible for keeping me in line with the logicist approach to actions and change; and Alex and David have taught me that counterfactuals are simpler than the name may imply.

My academic and professional colleagues have been very generous with their time and ideas as I began ploughing the peaceful territories of statistics, economics, epidemiology, philosophy, and the social sciences. My mentors–listeners in statistics have been Phil Dawid, Steffen Lauritzen, Don Rubin, Art Dempster, David Freedman, and David Cox. In economics, I have benefited from many discussions with John Aldrich, Kevin Hoover, James Heckman, Ed ~~Learner~~ Leamer, and Herbert Simon. My forays into epidemiology resulted in a most fortunate and productive collaboration with Sander Greenland and James Robins. Philosophical debates with James Woodward, Nancy Cartwright, Brian Skyrms, Clark Glymour, and Peter Spirtes have sharpened my thinking of causality in and outside philosophy. Finally, in artificial intelligence, I have benefited from discussions with and the encouragement of Nils Nilsson, Ray Reiter, Don Michie, Joe Halpern, and David Heckerman.

The National Science Foundation deserves acknowledgment for consistently and faithfully sponsoring the research that led to these results, with special thanks to H. Moraff, Y. T. Chien, and Larry Reeker. Other sponsors include Abraham Waksman of the Air Force Office of Scientific Research, Michael Shneier of the Office of Naval Research, the California MICRO Program, Northrop Corporation, Rockwell International, Hewlett-Packard, and Microsoft.

I would like to thank Academic Press and Morgan Kaufmann Publishers for their kind permission to reprint selected portions of previously published material. Chapter 3 includes material reprinted from *Biometrika,* vol. 82, Judea Pearl, "Causal Diagrams for Empirical Research," pp. 669–710, Copyright 1995, with permission from Oxford University Press. Chapter 5 includes material reprinted from *Sociological Methods and Research,* vol. 27, Judea Pearl, "Graphs, Causality, and Structural Equation Models," pp. 226–84, Copyright 1998, with permission from Sage Publications, Inc. Chapter 7 includes material reprinted from *Foundations of Science,* vol. 1, David Galles and Judea Pearl, "An Axiomatic Characterization of Causal Counterfactuals," pp. 151–82, Copyright 1998, with permission from Kluwer Academic Publishers. Chapter 7 also includes material reprinted from *Artificial Intelligence,* vol. 97, David Galles and Judea Pearl, "Axioms

(cf. equation (1.23)). Of special importance is the expectation of the product $g(X, Y) = (X - E(X))(Y - E(Y))$, which is known as the *covariance* of X and Y,

$$\sigma_{XY} \triangleq E\,[(X - E(X))(Y - E(Y))],$$

and which is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and the *regression coefficient*

$$r_{XY} \triangleq \rho_{XY} \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_Y^2}.$$

The *conditional* variance, covariance, and correlation coefficient, given $Z = z$, are defined in a similar manner, using the conditional distribution $P(x, y \mid z)$ in taking expectations. In particular, the *conditional correlation coefficient*, given $Z = z$, is defined as

$$\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z}\,\sigma_{Y|z}}. \tag{1.24}$$

Additional properties, specific to normal distributions, will be reviewed in Chapter 5 (Section 5.2.1).

The foregoing definitions apply to discrete random variables – that is, variables that take on finite or denumerable sets of values on the real line. The treatment of expectation and correlation is more often applied to continuous random variables, which are characterized by a *density function* $f(x)$ defined as follows:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

for any two real numbers a and b with $a < b$. If X is discrete, then $f(x)$ coincides with the probability function $P(x)$, once we interpret the integral through the translation

$$\int_{-\infty}^{\infty} f(x)dx \Longleftrightarrow \sum_x P(x). \tag{1.25}$$

Readers accustomed to continuous analysis should bear this translation in mind whenever summation is used in this book. For example, the expected value of a continuous random variable X can be obtained from (1.21), to read

$$E(X) = \int_{-\infty}^{\infty} xf\,(x)\,dx,$$

with analogous translations for the variance, correlation, and so forth.

We now turn to define *conditional independence* relationships among variables, a central notion in causal modelling.

1.1.5 Conditional Independence and Graphoids

Definition 1.1.2 (Conditional Independence)
Let $V = \{V_1, V_2, \ldots\}$ be a finite set of variables. Let $P(\cdot)$ be a joint probability function over the variables in V, and let X, Y, Z stand for any three subsets of variables in V. The sets X and Y are said to be conditionally independent *given Z if*

$$P(x \mid y, z) = P(x \mid z) \quad \textit{whenever} \quad P(y, z) > 0. \tag{1.26}$$

In words, learning the value of Y does not provide additional information about X, once we know Z. (Metaphorically, Z "screens off" X from Y.)

Equation (1.26) is a terse way of saying the following: For any configuration x of the variables in the set X and for any configurations y and z of the variables in Y and Z satisfying $P(Y = y, Z = z) > 0$, we have

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z). \tag{1.27}$$

We will use Dawid's (1979) notation $(X \perp\!\!\!\perp Y \mid Z)_P$ or simply $(X \perp\!\!\!\perp Y \mid Z)$ to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z) \tag{1.28}$$

for all values x, y, z such that $P(y, z) > 0$. Unconditional independence (also called *marginal independence*) will be denoted by $(X \perp\!\!\!\perp Y \mid \emptyset)$; that is,

$$(X \perp\!\!\!\perp Y \mid \emptyset) \text{ iff } P(x \mid y) = P(x) \qquad \text{whenever} \quad P(y) > 0 \tag{1.29}$$

("iff" is shorthand for "if and only if"). Note that $(X \perp\!\!\!\perp Y \mid Z)$ implies the conditional independence of all pairs of variables $V_i \in X$ and $V_j \in Y$, but the converse is not necessarily true.

The following is a (partial) list of properties satisfied by the conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$. (We use YW to abbreviate $Y \cup W$.)

Symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$.

Decomposition: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)$.

Weak union: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid ZW)$.

Contraction: $(X \perp\!\!\!\perp Y \mid Z)\ \&\ (X \perp\!\!\!\perp W \mid ZY) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

Intersection: $(X \perp\!\!\!\perp W \mid ZY)\ \&\ (X \perp\!\!\!\perp Y \mid ZW) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

(Intersection is valid in strictly positive probability distributions.)

The proof of these properties can be derived by elementary means from (1.28) and the basic axioms of probability theory.[4] These properties were called *graphoid axioms* by

[4] These properties were first introduced by Dawid (1979) and Spohn (1980) in a slightly different form, and were independently proposed by Pearl and Paz (1987) to characterize the relationships between graphs and informational relevance. Geiger and Pearl (1993) present an in-depth analysis.

boldface $\boldsymbol{P}$

Should be an empty set symbol - please match font used on page 11's "empty set" character; line before (1.29).

to constants x.[10] *Denote by* $\boldsymbol{P}_*$ *the set of all interventional distributions* $P_x(v), X \subseteq V$, *including* $P(v)$, *which represents no intervention (i.e.,* $X = \emptyset$*). A DAG* G *is said to be a* causal Bayesian network *compatible with* $\boldsymbol{P}_*$ *if and only if the following three conditions hold for every* $P_x \in \boldsymbol{P}_*$:

(i) $P_x(v)$ *is Markov relative to* G;

(ii) $P_x(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$;

(iii) $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$, i.e., each $P(v_i \mid pa_i)$ *remains invariant to interventions not involving* V_i.

Definition 1.3.1 imposes constraints on the interventional space $\boldsymbol{P}_*$ that permit us to encode this vast space economically, in the form of a single Bayesian network G. These constraints enable us to compute the distribution $P_x(v)$ resulting from any intervention $do(X = x)$ as a *truncated factorization*

$$P_x(v) = \prod_{\{i \mid V_i \notin X\}} P(v_i \mid pa_i) \qquad \text{for all } v \text{ consistent with } x, \tag{1.37}$$

which follows from Definition 1.3.1 and justifies the family deletion procedure on G, as in (1.36). It is not hard to show that, whenever G is a causal Bayes network with respect to $\boldsymbol{P}_*$, the following two properties must hold.

Property 1
For all i,

$$P(v_i \mid pa_i) = P_{pa_i}(v_i). \tag{1.38}$$

Property 2
For all i *and for every subset* S *of variables disjoint of* $\{V_i, PA_i\}$, *we have*

$$P_{pa_i, s}(v_i) = P_{pa_i}(v_i). \tag{1.39}$$

Property 1 renders every parent set PA_i *exogenous* relative to its child V_i, ensuring that the conditional probability $P(v_i \mid pa_i)$ coincides with the effect (on V_i) of setting PA_i to pa_i by external control. Property 2 expresses the notion of invariance; once we control its direct causes PA_i, no other interventions will affect the probability of V_i.

1.3.2 Causal Relationships and Their Stability

This mechanism-based conception of interventions provides a semantical basis for notions such as "causal effects" or "causal influence," to be defined formally and analyzed in Chapters 3 and 4. For example, to test whether a variable X_i has a causal influence on another variable X_j, we compute (using the truncated factorization formula of (1.37)) the (marginal) distribution of X_j under the actions $do(X_i = x_i)$ – namely, $P_{x_i}(x_j)$ for all

[10] The notation $P_x(v)$ will be replaced in subsequent chapters with $P(v \mid do(x))$ and $P(v \mid \hat{x})$ to facilitate algebraic manipulations.

Finally, certain concepts that are ubiquitous in human discourse can be defined only in the Laplacian framework. We shall see, for example, that such simple concepts as "the probability that event B occured *because* of event A" and "the probability that event B would have been *different* if it were not for event A" cannot be defined in terms of purely stochastic models. These so-called *counterfactual* concepts will require a synthesis of the deterministic and probabilistic components embodied in the Laplacian model.

1.4.1 Structural Equations

In its general form, a functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{1.40}$$

where pa_i (connoting *parents*) stands for the set of variables that directly determine the value of X_i and where the U_i represent errors (or "disturbances") due to omitted factors. Equation (1.40) is a nonlinear, nonparametric generalization of the linear structural equation models (SEMs)

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, \ldots, n, \tag{1.41}$$

which have become a standard tool in economics and social science (see Chapter 5 for a detailed exposition of this enterprise). In linear models, pa_i corresponds to those variables on the r.h.s. of (1.41) that have nonzero coefficients.

The interpretation of the functional relationship in (1.40) is the standard interpretation that functions carry in physics and the natural sciences; it is a recipe, a strategy, or a *law* specifying what value nature would assign to X_i in response to every possible value combination that (PA_i, U_i) might take on. A set of equations in the form of (1.40) and in which each equation represents an autonomous mechanism is called a *structural model*; if each variable has a distinct equation in which it appears on the left-hand side (called the *dependent* variable), then the model is called a *structural causal model* or a *causal model* for short.[13] ~~Mathematically, the distinction between structural and algebraic equations is that any subset of structural equations is, in itself, a valid structural model – one that represents conditions under some set of interventions.~~

To illustrate, Figure 1.5 depicts a canonical econometric model relating price and demand through the equations

$$q = b_1 p + d_1 i + u_1, \tag{1.42}$$

$$p = b_2 q + d_2 w + u_2, \tag{1.43}$$

where Q is the quantity of household demand for a product A, P is the unit price of product A, I is household income, W is the wage rate for producing product A, and U_1 and

cannot be ignored when the meaning of the concept is in question. Indeed, compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation, and the proper incorporation of background information into statistical studies likewise relies on accurate interpretation of causal judgment.

[13] Formal treatment of causal models, structural equations, and error terms are given in Chapter 5 (Section 5.4.1) and Chapter 7 (Sections 7.1 and 7.2.5).

Mathematically, the distinction between structural and algebraic equations is that the former change meaning under solution-preserving algebraic operations (e.g., moving terms from one side of an equation to the other.)

subject to falsification tests in the form of inequalities on conditional probabilities (Pearl 1995b). Still, such constraints permit the testing of merely a small fraction of the causal assumptions embodied in the diagrams; the bulk of those assumptions must be substantiated from domain knowledge as obtained from either theoretical considerations (e.g., that falling barometers do not cause rain) or related experimental studies. For example, the experimental study of Moertel et al. (1985), which refuted the hypothesis that vitamin C is effective against cancer, can be used as a substantive assumption in observational studies involving vitamin C and cancer patients; it would be represented as a missing link (between vitamin C and cancer) in the associated diagram. In summary, the primary use of the methods described in this chapter lies not in testing causal assumptions but in providing an effective language for making those assumptions precise and explicit. Assumptions can thereby be isolated for deliberation or experimentation and then (once validated) be integrated with statistical data to yield quantitative estimates of causal effects.

An important issue that will be considered only briefly in this book (see Section 8.5) is sampling variability. The mathematical derivation of causal effect estimands should be considered a first step toward supplementing these estimands with confidence intervals and significance levels, as in traditional analysis of controlled experiments. We should remark, though, that having obtained nonparametric estimands for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (3.28) can be converted to the product $E(Y \mid \hat{x}) = r_{ZX} r_{YZ \cdot X} x$, where $r_{YZ \cdot X}$ is the standardized regression coefficient (Section 5.3.1); the estimation problem then reduces to that of estimating regression coefficients (e.g., by least squares). More sophisticated estimation techniques can be found in Rosenbaum and Rubin (1983), Robins (1989, sec. 17), and Robins et al. (1992, pp. 331–3). For example, the "propensity score" method of Rosenbaum and Rubin (1983) was found useful when the dimensionality of the adjusted covariates is high (Section 11.3.5). Robins (1999) shows that, rather than estimating individual factors in the adjustment formula of (3.19), it is often more advantageous to use $P(y \mid \hat{x}) = \sum_z \frac{P(x, y, z)}{P(x \mid z)}$, where the preintervention distribution remains unfactorized. One can then separately estimate the denominator $P(x \mid z)$, weigh individual samples by the inverse of this estimate, and treat the weighted samples as if they were drawn at random from the postintervention distribution $P(y \mid \hat{x})$. Postintervention parameters, such as $\frac{\partial}{\partial x} E(Y \mid \hat{x})$, can then be estimated by ordinary least squares. This method is especially advantageous in longitudinal studies with time-varying covariates, as in the problems discussed in Sections 3.2.3 (see (3.18)) and 4.4.3.

, called *inverse probability weighting*,

Several extensions of the methods proposed in this chapter are noteworthy. First, the identification analysis for atomic interventions can be generalized to complex time-varying policies in which a set X of controlled variables is made to respond in a specified way to some set Z of covariates via functional or stochastic strategies, as in Sections 3.2.3 and 4.4.3. In Chapter 4 (Section 4.4.3) it is shown that identifying the effect of such policies requires a sequence of back-door conditions in the associated diagram.

A second extension concerns the use of the intervention calculus (Theorem 3.4.1) 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 3.2.1) still carries over to nonrecursive systems (Strotz and Wold

the temporally ordered and potentially manipulable treatment variables of interest. The causal effect of $X = x$ on Y was expressed as the probability

$$P(y \mid g = x) \triangleq P\{Y(x) = y\},$$

where the counterfactual variable $Y(x)$ stands for the value that outcome variables Y would take had the treatment variables X been x.

Robins showed that $P(y \mid g = x)$ is identified from the distribution $P(v)$ if each component X_k of X is "assigned at random, given the past," a notion explicated as follows in (Robins, 1995). Let L_k be the variables occurring between X_{k-1} and X_k, with L_1 being the variables preceding X_1, Write $\overline{L}_k = (L_1, \ldots, L_k)$, $L = \overline{L}_K$, and $\overline{X}_k = (X_1, \ldots, X_k)$, and define $\overline{X}_0$, $\overline{L}_0$, $\overline{V}_0$ to be identically zero. The treatment $X_k = x_k$ is said to be *assigned at random, given the past*, if the following relation holds:

$$(Y(x) \perp\!\!\!\perp X_k \mid \overline{L}_k, \overline{X}_{k-1} = \overline{x}_{k-1}). \tag{3.62}$$

Robins further proved that, if (3.62) holds for every k, then the causal effect is given by

$$P(y \mid g = x) = \sum_{\overline{l}_K} P(y \mid \overline{l}_K, \overline{x}_K) \prod_{k=1}^{K} P(l_k \mid \overline{l}_{k-1}, \overline{x}_{k-1}), \tag{3.63}$$

sub cap K

an expression he called the "G-computation algorithm formula." This expression can be derived by applying condition (3.62) iteratively, as in the derivation of (3.54). If X is univariate, then (3.63) reduces to the standard adjustment formula

$$P(y \mid g = x) = \sum_{l_1} P(y \mid x, l_1)\, P(l_1),$$

paralleling (3.54). Likewise, in the special structure of Figure 3.3, (3.63) reduces to (3.18).

To place this result in the context of our analysis in this chapter, we need to focus attention on condition (3.62), which facilitated Robins's derivation of (3.63), and ask whether this formal counterfactual independency can be given a meaningful graphical interpretation. The answer will be given in Chapter 4 (Theorem 4.4.1), where we derive a graphical condition for identifying the effect of a plan, i.e., a sequential set of actions. Alternatively, (3.62) can be obtained by applying the translation rule of (3.56) to graphs with no confounding arcs between X_k and $\{PA_k\}$. We note, however, that the implication goes only one way; Robin's condition (3.62) does not imply the graph structure, because the translation of (3.56) asserts joint independencies among counterfactuals (see footnote 14, page 101), which is not required in (3.62).

The structural analysis introduced in this chapter supports and generalizes Robins's result from a new theoretical perspective. First, on the technical front, this analysis offers systematic ways of managing models where Robins's starting assumption (3.62) is inapplicable. Examples are Figures 3.8(d)–(g).

Second, on the conceptual front, the structural framework represents a fundamental shift from the vocabulary of counterfactual independencies, to the vocabulary of

Phil showed special courage in printing my paper in *Biometrika* (Pearl 1995a), the journal founded by causality's worst adversary – Karl Pearson.

Postscript for the Second Edition

Complete identification results

A key identification condition, which generalizes all the criteria established in this chapter, has been derived by Jin Tian. It reads:

Theorem 3.6.1 (Tian and Pearl, 2002a)
A sufficient condition for identifying the causal effect $P(y \mid do(x))$ *is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between* X *and any of its children.*[15]

Remarkably, the theorem asserts that, as long as every child of X (on the pathways to Y) is not reachable from X via a bi-directed path, then, regardless of how complicated the graph, the causal effect $P(y \mid do(x))$ is identifiable. All identification criteria discussed in this chapter are special cases of the one defined in this theorem. For example, in Figure 3.5 $P(y \mid do(x))$ can be identified because the two paths from X to Z (the only child of X) are not bi-directed. In Figure 3.7, on the other hand, there is a path from X to Z_1 traversing only bi-directed arcs, thus violating the condition of Theorem 3.6.1, and $P(y \mid do(x))$ is not identifiable.

Note that all graphs in Figure 3.8 and none of those in Figure 3.9 satisfy the condition above. Tian and Pearl (2002a) further showed that the condition is both sufficient and necessary for the identification of $P(v \mid do(x))$, where V includes all variables except X. A necessary and sufficient condition for identifying $P(w \mid do(z))$, with W and Z two arbitrary sets, was established by Shpitser and Pearl (2006b). Subsequently, a complete graphical criterion was established for determining the identifiability of *conditional* interventional distributions, namely, expressions of the type $P(y \mid do(x), z)$ where X, Y, and Z are arbitrary sets of variables (Shpitser and Pearl 2006a).

These results constitute a complete characterization of causal effects in graphical models. They provide us with polynomial time algorithms for determining whether an arbitrary quantity invoking the $do(x)$ operator is identified in a given semi-Markovian model and, if so, what the estimand of that quantity is. Remarkably, one corollary of these results also states that the *do*-calculus is complete, namely, a quantity $Q = P(y \mid do(x), z)$ is identified if and only if it can be reduced to a *do*-free expression using the three rules of Theorem 3.4.1.[16] Pearl (2012c) describes new applications of *do*-calculus. Tian and Shpitser (2010) provide a comprehensive summary of these results.

Applications and Critics

Gentle introductions to the concepts developed in this chapter are given in (Pearl 2003c) and (Pearl 2008). Applications of causal graphs in epidemiology are reported in Robins

[15] Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.
[16] This was independently established by Huang and Valtorta (2006).

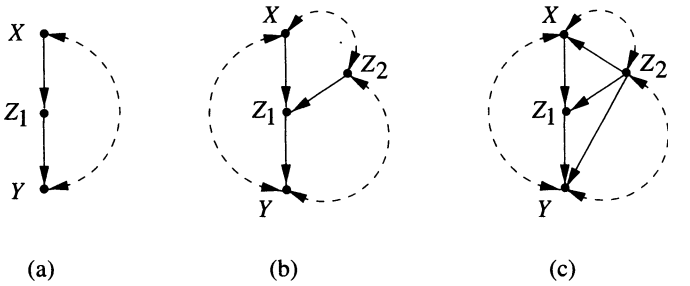

Figure 4.2 Condition 4 of Theorem 4.3.1. In (a), Z_1 blocks all directed paths from X to Y, and the empty set blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G. In (b) and (c), Z_1 blocks all directed paths from X to Y, and Z_2 blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G.

summing over X and so derive $\sum_{x'} P(y \mid \hat{z}_1, z_2, x') P(x' \mid \hat{z}_1, z_2)$. Now we can rewrite $P(y \mid \hat{z}_1, z_2, x')$ as $P(y \mid z_1, z_2, x')$ using Rule 2. The $P(x' \mid \hat{z}_1, z_2)$ term can be rewritten as $P(x' \mid z_2)$ using Rule 3, since Z_1 is a child of X and the graph is acyclic. The query can therefore be rewritten as $\sum_{z_1, z_2} \sum_{x'} P(y \mid z_1, z_2, x') P(x' \mid z_2) P(z_1, z_2 \mid \hat{x})$, and we have $P(z_1, z_2 \mid \hat{x}) = P(z_2 \mid \hat{x}) P(z_1 \mid \hat{x}, z_2)$. Since Z_2 consists of nondescendants of X, we can rewrite $P(z_2 \mid \hat{x})$ as $P(z_2)$ using Rule 3. Since Z_2 blocks all back-door paths from X to Z_1. we can rewrite $P(z_1 \mid \hat{x}, z_2)$ as $P(z_1 \mid x, z_2)$ using Rule 2. The entire query can thus be rewritten as $\sum_{z_1, z_2} \sum_{x'} P(y \mid z_1, z_2, x') P(x' \mid z_2) P(z_1 \mid x, z_2) P(z_2)$. See examples in Figure 4.2. □

Theorem 4.3.2 (Retracted)

~~*At leas*~~
~~*is, if a*~~
~~*sequer*~~

~~A proc~~

An earlier edition of this book contained a necessary condition for identifying $P(y|x)$, which was later found to be incomplete. A correct necessary and sufficient condition is given in Tian and Pearl (2003), which was proven to be complete by Shpitser and Pearl (2006b) and Huang and Valtorta (2006b).

4.3.2 Remarks on Efficiency

In implementing Theorem 4.3.1 as a systematic method for determining identifiability, Conditions 3 and 4 would seem to require exhaustive search. In order to prove that Condition 3 does not hold, for instance, we need to prove that no such blocking set B can exist. Fortunately, the following theorems allow us to significantly prune the search space so as to render the test tractable.

Theorem 4.3.3

If $P(b_i \mid \hat{x})$ is identifiable for one minimal set B_i, then $P(b_j \mid \hat{x})$ is identifiable for any other minimal set B_j.

Theorem 4.3.3 allows us to test Condition 3 with a single minimal blocking set B. If B meets the requirements of Condition 3, then the query is identifiable; otherwise, Condition 3 cannot be satisfied. In proving this theorem, we use the following lemma.

Two other features of Figure 4.4 are worth noting. First, the quantity $P(y \mid \hat{x}_1, \hat{x}_2)$ cannot be computed if we treat the control variables X_1 and X_2 as a single compound variable X. The graph corresponding to such compounding would depict X as connected to Y by both an arrow and a curved arc (through U) and thus would form a bow pattern (see Figure 3.9), which is indicative of nonidentifiability. Second, the causal effect $P(y \mid \hat{x}_1)$ in isolation is not identifiable because U_1 creates a bow pattern around the link $X \rightarrow Z$, which lies on a directed path from X to Y (see the discussion in Section 3.5).

The feature that facilitates the identifiability of $P(y \mid \hat{x}_1, \hat{x}_2)$ is the identifiability of $P(y \mid x_1, z, \hat{x}_2)$ – the causal effect of the action $do(X_2 = x_2)$ alone, conditioned on the observations available at the time of this action. This can be verified using the back-door criterion, observing that $\{X_1, Z\}$ blocks all back-door paths between X_2 and Y. Thus, the identifiability of $P(y \mid \hat{x}_1, \hat{x}_2)$ can be readily proven by writing

$$P(y \mid \hat{x}_1, \hat{x}_2) = P(y \mid x_1, \hat{x}_2) \tag{4.1}$$

$$= \sum_z P(y \mid z, x_1, \hat{x}_2) P(z \mid x_1) \tag{4.2}$$

$$= \sum_z P(y \mid z, x_1, x_2) P(z \mid x_1), \tag{4.3}$$

where (4.1) and (4.3) follow from Rule 2, and (4.2) follows from Rule 3. The subgraphs that permit the application of these rules are shown in Figure 4.5 (in Section 4.4.3).

This derivation also highlights how conditional plans can be evaluated. Assume we wish to evaluate the effect of the plan $\{do(X_1 = x_1), do(X_2 = g(x_1, z))\}$. Following the analysis of Section 4.2, assuming $U_1 = \{0\}$, we write

$$P(y \mid do(X_1 = x_1), do(X_2 = g(x_1, z))) = P(y \mid x_1, do(X_2 = g(x_1, z)))$$

$$= \sum_z P(y \mid z, x_1, do(X_2 = g(x_1, z))) P(z \mid x_1)$$

$$= \sum_z P(y \mid z, x_1, x_2) P(z \mid x_1)|_{x_2 = g(x_1, z)}.$$

Again, the identifiability of this conditional plan rests on the identifiability of the expression $P(y \mid z, x_1, \hat{x}_2)$, which reduces to $P(y \mid z, x_1, x_2)$ because $\{X_1, Z\}$ blocks all back-door paths between X_2 and Y. (See also Section 11.4.1.)

The criterion developed in the next section will enable us to recognize in general, by graphical means, whether a proposed plan can be evaluated from the joint distribution on the observables and, if so, to identify which covariates should be measured and how they should be adjusted.

4.4.2 Plan Identification: Notation and Assumptions

Our starting point is a knowledge specification scheme in the form of a causal diagram, like the one shown in Figure 4.4, that provides a qualitative summary of the analyst's understanding of the relevant data-generating processes.[5]

[5] An alternative specification scheme using counterfactual dependencies was used in Robins (1986, 1987), as described in Section 3.6.4.

Proof of Theorem 4.4.1

The proof given here is based on the inference rules of *do*-calculus (Theorem 3.4.1), which facilitate the reduction of causal effect formulas to hat-free expressions. An alternative proof, using latent variable elimination, is given in Pearl and Robins (1995).

Step 1. The condition $Z_k \subseteq N_k$ implies $Z_k \subseteq N_j$ for all $j \geq k$. Therefore, we have

$$P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}, \hat{x}_k, \hat{x}_{k+1}, \ldots, \hat{x}_n) = P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}).$$

This is so because no node in $\{Z_1, \ldots, Z_k, X_1, \ldots, X_{k-1}\}$ can be a descendant of any node in $\{X_k, \ldots, X_n\}$. Hence, Rule 3 allows us to delete the hat variables from the expression.

Step 2. The condition in (4.5) permits us to invoke Rule 2 and write:

$$P(y \mid z_1, \ldots, z_k, x_1, \ldots, x_{k-1}, \hat{x}_k, \hat{x}_{k+1}, \ldots, \hat{x}_n) = P(y \mid z_1, \ldots, z_k, x_1, \ldots, x_{k-1}, x_k, \hat{x}_{k+1}, \ldots, \hat{x}_n).$$

Thus, we have

$$\begin{aligned}
&P(y \mid \hat{x}_1, \ldots, \hat{x}_n) \\
&= \sum_{z_1} P(y \mid z_1, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1 \mid \hat{x}_1, \ldots, \hat{x}_n) \\
&= \sum_{z_1} P(y \mid z_1, x_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1) \\
&= \sum_{z_2} \sum_{z_1} P(y \mid z_1, z_2, x_1, \hat{x}_2, \ldots, \hat{x}_n) P(z_1) P(z_2) \mid z_1, x_1, \hat{x}_2, \ldots, \hat{x}_n) \\
&= \sum_{z_2} \sum_{z_1} P(y \mid z_1, z_2, x_1, x_2, \hat{x}_3, \ldots, \hat{x}_n) P(z_1) P(z_2) \mid z_1, x_1) \\
&\vdots \\
&= \sum_{z_n} \cdots \sum_{z_2} \sum_{z_1} P(y \mid z_1, \ldots, z_n, x_1, \ldots, x_n) \\
&\quad \times P(z_1) P(z_2 \mid z_1, x_1) \cdots P(z_n \mid z_1, x_1, z_2, x_2, \ldots, z_{n-1}, x_{n-1}) \\
&= \sum_{z_1, \ldots, z_n} P(y \mid z_1, \ldots, z_n, x_1, \ldots, x_n) \prod_{k=1}^{n} P(z_k \mid z_1, \ldots, z_{k-1}, x_1, \ldots, x_{k-1}). \qquad \square
\end{aligned}$$

Definition 4.4.2 (Admissible Sequence and *G*-Identifiability)

Any sequence $Z_1, \ldots, Z_n$ *of covariates satisfying the conditions in* (4.4)–(4.5) *will be called* admissible, *and any expression* $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ *that is identifiable by the criterion of Theorem 4.4.1 will be called* G-identifiable.[7]

, in graphical terms, the

[7] Note that admissibility (4.5) requires that ~~each~~ subsequence $X_1, \ldots, X_{k-1}, Z_1, \ldots, Z_k$ blocks every "action-avoiding" back-door path from X_k to Y (see ~~page 1037~~ (3.62*), page 352).

$$P(\text{admission} \mid \widehat{\text{male}}, \widehat{\text{dept}}) - P(\text{admission} \mid \widehat{\text{female}}, \widehat{\text{dept}})$$

with some average of this difference over all departments. This average should measure the increase in admission rate in a hypothetical experiment in which we instruct all female candidates to retain their department preferences but change their gender identification (on the application form) from female to male.

Conceptually, we can define the average direct effect $DE_{x,x'}(Y)$ as the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they would have obtained under $do(x)$. This hypothetical change, which Robins and Greenland (1992) called "pure" and Pearl (2001c) called "natural," is precisely what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)).

Using the parenthetical notation of equation 3.51, Pearl (2001c) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E[(Y(x', Z(x))) - E(Y(x)]. \tag{4.11}$$

Here, Z represents all parents of Y excluding X, and the expression $Y(x', Z(x))$ represents the value that Y would attain under the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. We see that $DE_{x,x'}(Y)$, the natural direct effect of the transition from x to x', involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified, even with the help of ideal, controlled experiments (see Robins and Greenland 1992 and Section 7.1 for intuitive explanation). Pearl (2001c) has nevertheless shown that, if certain assumptions of "no confounding" are deemed valid,[9] the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y \mid do(x', z)) - E(Y \mid do(x, z))]\, P(z \mid do(x)). \tag{4.12}$$

The intuition is simple; the natural direct effect is the weighted average of controlled direct effects, using the causal effect $P(z \mid do(x))$ as a weighing function. Under such assumptions, the sequential back-door criteria developed in Section 4.4 for identifying control-specific plans, $P(y \mid \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$, become applicable.

In particular, ~~expression (4.12) is both valid and~~ identifiable in Markovian models ~~where all *do*-operators can be eliminated using Corollary 3.2.6; for example,~~ through:

$$DE_{x,x'}(Y) = \sum_{zw} [E(Y \mid x', z, w) - E(Y \mid x, z, w)]\, P(z \mid x, w)\, P(w)$$

~~$$P(z \mid do(x)) = \sum_t P(z \mid x, pa_X = t)\, P(pa_X = t)$$~~ (4.13)

where W satisfies the back-door criterion relative to both $X \to Z$ and $(X,Z) \to Y$. (See Pearl (2001c; 2012b,c) and Shpitser and VanderWeele (2011).)

[9] One sufficient condition is that $Z(x) \perp\!\!\!\perp Y(x', z) \mid W$ holds for some set W of measured covariates. See details and graphical criteria in Pearl (2001c, ~~2005a~~) and in Petersen et al. (2006).

4.5.5 Indirect Effects

Remarkably, the definition of the natural direct effect (4.11) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the $do(x)$ operator, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The natural indirect effect, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads (Pearl 2001c):

$$IE_{x,x'}(Y) \triangleq E[(Y(x, Z(x'))) - E(Y(x))], \tag{4.14}$$

which is almost identical to the direct effect (equation (4.11)) save for exchanging x and x'.

Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \triangleq E(Y(x) - Y(x')) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{4.15}$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{4.16}$$

Since each term above is based on an independent operational definition, this quality constitutes a formal justification for the additive formula.

Note that the indirect effect has clear policy-making implications. For example: in a hiring discrimination environment, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally – say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent. See (Pearl 2001c, 2012a) for more examples.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of X on Y through a selected set of paths (Avin et al. 2005).

Note that in all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001c) has suggested therefore that signal sensing is more fundamental to the notion of causation than manipulation; the latter being but a crude way of stimulating the former in experimental setup. (See Section 11.4.5.)

It is remarkable that counterfactual quantities like DE and IE that could not be expressed in terms of $do(x)$ operators, and appear therefore void of empirical content, can, under certain conditions, be estimated from empirical studies. A general analysis of those conditions is given in Shpitser and Pearl (2007).

We shall see additional examples of this "marvel of formal analysis" in Chapters 7, 9, and 11. It constitutes an unassailable argument in defense of counterfactual analysis, as expressed in Pearl (2000) against the stance of Dawid (2000).

This model is as compact as (5.7)–(5.9) and is covariance equivalent to M with respect to the observed variables X, Y, Z. Upon setting $\alpha' = \alpha, \beta' = \beta$, and $\delta = \gamma$, model M' will yield the same probabilistic predictions as those of the model of (5.7)–(5.9). Still, when viewed as data-generating mechanisms, the two models are not equivalent. Each tells a different story about the processes generating X, Y, and Z, so naturally their predictions differ concerning the changes that would result from subjecting these processes to external interventions.

5.3.3 Causal Effects: The Interventional Interpretation of Structural Equation Models

The differences between models M and M' illustrate precisely where the structural reading of simultaneous equation models comes into play, and why even causally shy researchers consider structural parameters more "meaningful" than covariances and other statistical parameters. Model M', defined by (5.12)–(5.14), regards X as a direct participant in the process that determines the value of Y, whereas model M, defined by (5.7)–(5.9), views X as an indirect factor whose effect on Y is mediated by Z. This difference is not manifested in the data itself but rather in the way the data would change in response to outside interventions. For example, suppose we wish to predict the expectation of Y after we intervene and fix the value of X to some constant x; this is denoted $E(Y \mid do(X = x))$. After $X = x$ is substituted into (5.13) and (5.14), model M' yields

$$E[Y \mid do(X = x)] = E[\beta'\alpha' x + \beta'\varepsilon_2 + \delta x + \varepsilon_3] \tag{5.15}$$
$$= (\beta'\alpha' + \delta)x; \tag{5.16}$$

and model M yields

$$E[Y \mid do(X = x)] = E[\beta\alpha x + \beta\varepsilon_2 + \gamma u + \varepsilon_3] \tag{5.17}$$
$$= \beta\alpha x. \tag{5.18}$$

Upon setting $\alpha' = \alpha, \beta' = \beta$, and $\delta = \gamma$ (as required for covariance equivalence; see (5.10) and (5.11)), we see clearly that the two models assign different magnitudes to the (total) causal effect of X on Y: model M predicts that a unit change in x will change $E(Y)$ by the amount $\beta\alpha$, whereas model M' puts this amount at $\beta\alpha + \delta$. δ (lowercase delta)

At this point, it is tempting to ask whether we should substitute $x - \varepsilon_1$ for u in (5.9) prior to taking expectations in (5.17). If we permit the substitution of (5.8) into (5.9), as we did in deriving (5.17), why not permit the substitution of (5.7) into (5.9) as well? After all (the argument runs), there is no harm in upholding a mathematical equality, $u = x - \varepsilon_1$, that the modeler deems valid. This argument is fallacious, however.[15] Structural equations are not meant to be treated as immutable mathematical equalities. Rather, they are meant to define a state of equilibrium – one that is *violated* when the equilibrium is perturbed by outside interventions. In fact, the power of structural equation models is

[15] Such arguments have led to Newcomb's paradox in the so-called evidential decision theory (see Section 4.1.1).

that they encode not only the initial equilibrium state but also the information necessary for determining which equations must be violated in order to account for a new state of equilibrium. For example, if the intervention consists merely of holding X constant at x, then the equation $x = u + \varepsilon_1$, which represents the preintervention process determining X, should be overruled and replaced with the equation $X = x$. The solution to the new set of equations then represents the new equilibrium. Thus, the essential characteristic of structural equations that sets them apart from ordinary mathematical equations is that the former stand not for one but for many sets of equations, each corresponding to a subset of equations taken from the original model. Every such subset represents some hypothetical physical reality that would prevail under a given intervention.

If we take the stand that the value of structural equations lies not in summarizing distribution functions but in encoding causal information for predicting the effects of policies (Haavelmo 1943; Marschak 1950; Simon 1953), it is natural to view such predictions as the proper generalization of structural coefficients. For example, the proper generalization of the coefficient β in the linear model M would be the answer to the control query, "What would be the change in the expected value of Y if we were to intervene and change the value of Z from z to $z + 1$?", which is different, of course, from the observational query, "What would be the difference in the expected value of Y if we were to *find* Z at level $z + 1$ instead of level z?" Observational queries, as we discussed in Chapter 1, can be answered directly from the joint distribution $P(x, y, z)$, while control queries require causal information as well. Structural equations encode this causal information in their syntax by treating the variable on the left-hand side of the equality sign as the effect and treating those on the right as causes. In Chapter 3 we distinguished between the two types of queries through the symbol $do(\cdot)$. For example, we wrote

$$E(Y \mid do(x)) \triangleq E\,[Y \mid do(X = x)] \tag{5.19}$$

for the controlled expectation and

$$E(Y \mid x) \triangleq E(Y \mid X = x) \tag{5.20}$$

for the standard conditional or observational expectation. That $E(Y \mid do(x))$ does not equal $E(Y \mid x)$ can easily be seen in the model of (5.7)–(5.9), where $E(Y \mid do(x)) = \alpha\beta x$ but $E(Y \mid x) = r_{YX}x = (\alpha\beta + \gamma)\, x$. Indeed, the passive observation $X = x$ should not violate any of the equations, and this is the justification for substituting both (5.7) and (5.8) into (5.9) before taking the expectation.

γ (lowercase gamma)

In linear models, the answers to questions of direct control are encoded in the path (or structural) coefficients, which can be used to derive the total effect of any variable on another. For example, the value of $E(Y \mid do(x))$ in the model defined by (5.7)–(5.9) is $\alpha\beta x$, that is, x times the product of the path coefficients along the path $X \rightarrow Z \rightarrow Y$. Computation of $E(Y \mid do(x))$ would be more complicated in the nonparametric case, even if we knew the functions f_1, f_2, and f_3. Nevertheless, this computation is well defined; it requires the solution (for the expectation of Y) of a modified set of equations in which f_1 is "wiped out" and X is replaced by the constant x:

$$z = f_2(x, \varepsilon_2), \tag{5.21}$$

$$y = f_3(z, u, \varepsilon_3). \tag{5.22}$$

"the concept of exogeneity rapidly evolved into a loose notion as a property of an observable variable being uncorrelated with an unobserved error," and Imbens (1997) readily agreed that this notion "is inadequate."[26]

These critics are hardly justified if we consider the precision and clarity with which structural errors can be defined when using the proper notation (e.g., (5.25)). When applied to structural errors, the standard error-based criterion of exogeneity coincides formally with that of (5.30), as can be verified using the back-door test of Theorem 5.3.2 (with $Z = \emptyset$). Consequently, the standard definition conveys the same information as that embodied in more complicated and less communicable definitions of exogeneity. I am therefore convinced that the standard definition will eventually regain the acceptance and respectability that it has always deserved.

Relationships between graphical and counterfactual definitions of exogeneity and instrumental variables will be discussed in Chapter 7 (Section 7.4.5).

5.5 CONCLUSION

Today the enterprise known as structural equation modeling is increasingly under fire. The founding fathers have retired, their teachings are forgotten, and practitioners, teachers, and researchers currently find the methodology they inherited difficult to either defend or supplant. Modern SEM textbooks are preoccupied with parameter estimation and rarely explicate the role that those parameters play in causal explanations or in policy analysis; examples dealing with the effects of interventions are conspicuously absent, for instance. Research in SEM now focuses almost exclusively on model fitting, while issues pertaining to the meaning and usage of SEM's models are subjects of confusion and controversy. Some of these confusions are reflected in the many questions that I have received from readers (Section 11.5), to whom I dedicated an "SEM Survival Kit" (Section 11.5.3) – a set of arguments for defending the causal reading of SEM and its scientific rationale.

I am thoroughly convinced that the contemporary crisis in SEM originates in the lack of a mathematical language for handling the causal information embedded in structural equations. Graphical models have provided such a language. They have thus helped us answer many of the unsettled questions that drive the current crisis:

1. Under what conditions can we give causal interpretation to structural coefficients?
2. What are the causal assumptions underlying a given structural equation model?
3. What are the statistical implications of any given structural equation model?
4. What is the operational meaning of a given structural coefficient?
5. What are the policy-making claims of any given structural equation model?
6. When is an equation not structural?

[26] Imbens prefers definitions in terms of experimental metaphors such as "random assignment assumption," fearing, perhaps, that "[t]ypically the researcher does not have a firm idea what these disturbances really represent" (Angrist et al. 1996, p. 446). I disagree; "random assignment" is a misleading metaphor, while "omitted factors" shines in clarity.

nebulous and potentially

unless the researcher first attains "a firm idea what these disturbances really represent,"

neither ensures unbiased effect estimates nor follows from the requirement of unbiasedness. After demonstrating, by examples, the absence of logical connections between the statistical and the causal notions of confounding, we will define a stronger notion of unbiasedness, called "stable" unbiasedness, relative to which a modified statistical criterion will be shown necessary and sufficient. The necessary part will then yield a practical test for stable unbiasedness that, remarkably, does not require knowledge of all potential confounders in a problem. Finally, we will argue that the prevailing practice of substituting statistical criteria for the effect-based definition of confounding is not entirely misguided, because stable unbiasedness is in fact (i) what investigators have been (and perhaps should be) aiming to achieve and (ii) what statistical criteria can test.

6.2.2 Causal and Associational Definitions

In order to facilitate the discussion, we shall first cast the causal and statistical definitions of no-confounding in mathematical forms.[11]

Definition 6.2.1 (No-Confounding; Causal Definition)
Let M be a causal model of the data-generating process – that is, a formal description of how the value of each observed variable is determined. Denote by $P(y \mid do(x))$ the probability of the response event $Y = y$ under the hypothetical intervention $X = x$, calculated according to M. We say that X and Y are not confounded in M if and only if

$$P(y \mid do(x)) = P(y \mid x) \quad \text{or} \quad P(x \mid do(y)) = P(x \mid y) \tag{6.10}$$

for all x and y in their respective domains, where $P(y \mid x)$ is the conditional probability generated by M. If (6.10) *holds, we say that $P(y \mid x)$ is unbiased.*

For the purpose of our discussion here, we take this causal definition as the meaning of the expression "no confounding." The probability $P(y \mid do(x))$ was defined in Chapter 3 (Definition 3.2.1, also abbreviated $P(y \mid \hat{x})$); it may be interpreted as the conditional probability $P^*(Y = y \mid X = x)$ corresponding to a controlled experiment in which X is randomized. We recall that this probability can be calculated from a causal model M either directly, by simulating the intervention $do(X = x)$, or (if $P(x, s) > 0$) via the adjustment formula (equation (3.19))

$$P(y \mid do(x)) = \sum_s P(y \mid x, s)\, P(s),$$

where S stands for any set of variables, observed as well as unobserved, that satisfy the back-door criterion (Definition 3.3.1). Equivalently, $P(y \mid do(x))$ can be written $P(Y(x) = y)$, where $Y(x)$ is the potential-outcome variable as defined in (3.51) or in

[11] For simplicity, we will limit our discussion to unadjusted confounding; extensions involving measurement of auxiliary variables are straightforward and can be obtained from Section 3.3. We also use the abbreviated expression "X and Y are not confounded," though "the effect of X on Y is not confounded" is more exact.

any sentence of the form $P(A \mid B) < p$, where A and B are Boolean expressions representing events. A *causal model*, naturally, should encode the truth values of sentences that deal with causal relationships; these include action sentences (e.g., "A will be true if we do B"), counterfactuals (e.g., "A would have been different were it not for B"), and plain causal utterances (e.g., "A may cause B" or "B occurred because of A"). Such sentences cannot be interpreted in standard propositional logic or probability calculus because they deal with changes that occur in the external world rather than with changes in our beliefs about a static world. Causal models encode and distinguish information about external changes through an explicit representation of the mechanisms that are altered in such changes.

Structural

Definition 7.1.1 (Causal Model)

A causal model *is a triple*

$$M = \langle U, V, F \rangle,$$

where:

(i) U *is a set of* background *variables,* (*also called* exogenous),[2] *that are determined by factors outside the model;*

(ii) V *is a set* $\{V_1, V_2, \ldots, V_n\}$ *of variables, called* endogenous, *that are determined by variables in the model – that is, variables in* $U \cup V$; *and*

(iii) F *is a set of functions* $\{f_1, f_2, \ldots, f_n\}$ *such that each* f_i *is a mapping from* (*the respective domains of*) $U_i \cup PA_i$ *to* V_i, *where* $U_i \subseteq U$ *and* $PA_i \subseteq V \backslash V_i$ *and the entire set* F *forms a mapping from* U *to* V. *In other words, each* f_i *in*

$$v_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n,$$

assigns a value to V_i *that depends on* (*the values of*) *a select set of variables in* $V \cup U$, *and the entire set* F *has a unique solution* $V(u)$.[3,4]

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable and the directed edges point from members of PA_i and U_i toward V_i. We call such a graph the *causal diagram* associated with M. This graph merely identifies the endogenous and background variables that have direct influence on each V_i; it does not specify the functional form of f_i. The convention of confining the parent set PA_i to variables in V stems from the fact that the background variables are often unobservable. In general, however, we can extend the parent sets to include observed variables in U.

---

[2] We will try to refrain from using the term "exogenous" in referring to background conditions, because this term has acquired more refined technical connotations (see Sections 5.4.3 and 7.4). The term "predetermined" is used in the econometric literature.

[3] The choice of PA_i (connoting *parents*) is not arbitrary, but expresses the modeller's understanding of which variables Nature must consult before deciding the value of V_i.

[4] Uniqueness is ensured in recursive (i.e., acyclic) systems. Halpern (1998) allows multiple solutions in nonrecursive systems.

utilization in policy predictions. Accordingly, we will demonstrate how to evaluate the following three queries.

1. What is the expected value of the demand Q if the price is *controlled at* $P = p_0$?
2. What is the expected value of the demand Q if the price is *reported to be* $P = p_0$?
3. Given that the current price is $P = p_0$, what would be the expected value of the demand Q if we *were to control* the price at $P = p_1$?

The reader should recognize these queries as representing (respectively) actions, predictions, and counterfactuals – our three-level hierarchy. The second query, representing prediction, is standard in the literature and can be answered directly from the covariance matrix without reference to causality, structure, or invariance. The first and third queries rest on the structural properties of the equations and, as expected, are not treated in the standard literature of structural equations.[10]

In order to answer the first query, we replace (7.10) with $p = p_0$, leaving

$$q = b_1 p + d_1 i + u_1, \tag{7.11}$$

$$p = p_0, \tag{7.12}$$

with the statistics of U_1 and I unaltered. The controlled demand is then $q = b_1 p_0 + d_1 i + u_1$, and its expected value (conditional on $I = i$) is given by

$$E[Q \mid do(P = p_0), i] = b_1 p_0 + d_1 i + E(U_1 \mid i). \tag{7.13}$$

Since U_1 is independent of I, the last term evaluates to

$$E(U_1 \mid i) = E(U_1) = E(Q) - b_1 E(P) - d_1 E(I)$$

and, substituted into (7.13), yields

$$E[Q \mid do(P = p_0), i] = E(Q) + b_1(p_0 - E(P)) + d_i (i - E(I)).$$

The answer to the second query is obtained by conditioning (7.9) on the current observation $\{P = p_0, I = i, W = w\}$ and taking the expectation,

$$E(Q \mid p_0, i, w) = b_1 p_0 + d_1 i + E(U_1 \mid p_0, i, w). \tag{7.14}$$

The computation of $E[U_1 \mid p_0, i, w]$ is a standard procedure once $\sum_{ij}$ is given (Whittaker 1990, p. 163). Note that, although U_1 was assumed to be independent of I and W, this independence no longer holds once $P = p_0$ is observed. Note also that (7.9) and (7.10)

[10] I have presented this example to well over a hundred econometrics students and faculty across the United States. Respondents had no problem answering question 2, one person was able to solve question 1, and none managed to answer question 3. Chapter 5 (Section 5.1) suggests an explanation, and Section 11.5.4 a more recent assessment based on Heckman and Vytlacil (2007).

Property 5 (Uniqueness)
For every variable X and set of variables Y,

$$X_y(u) = x \ \& \ X_y(u) = x' \Longrightarrow x = x'. \tag{7.23}$$

Definition 7.3.4 (Recursiveness)
Let X and Y be singleton variables in a model, and let $X \rightarrow Y$ stand for the inequality $Y_{xw}(u) \neq Y_w(u)$ for some values of x, w, and u. A model M is recursive if, for any sequence $X_1, X_2, \ldots, X_k$, we have

$$X_1 \rightarrow X_2, X_2 \rightarrow X_3, \ldots, X_{k-1} \rightarrow X_k \Longrightarrow X_k \not\rightarrow X_1. \tag{7.24}$$

Clearly, any model M for which the causal diagram $G(M)$ is acyclic must be recursive.

Theorem 7.3.5 (Recursive Completeness)
Composition, effectiveness, and recursiveness are complete (Galles and Pearl 1998; Halpern 1998).[15]

Theorem 7.3.6 (Completeness)
Composition, effectiveness, and reversibility are complete for all causal models (Halpern 1998).

The practical importance of soundness and completeness surfaces when we attempt to test whether a certain set of conditions is sufficient for the identifiability of some counterfactual quantity Q. Soundness, in this context, guarantees that if we symbolically manipulate Q using the three axioms and manage to reduce it to an expression that involves ordinary probabilities (free of counterfactual terms), then Q is identifiable (in the sense of Definition 3.2.3). Completeness guarantees the converse: if we do not succeed in reducing Q to a probabilistic expression, then Q is nonidentifiable – our three axioms are as powerful as can be.

The next section demonstrates a proof of identifiability that uses effectiveness and decomposition as ~~inference rules~~ axioms.

7.3.2 Causal Effects from Counterfactual Logic: An Example

We revisit the smoking–cancer example analyzed in Section 3.4.3. The model associated with this example is assumed to have the following structure (see Figure 7.5):

$$V = \{X \text{ (smoking)}, Y \text{ (lung cancer)}, Z \text{ (tar in lungs)}\},$$

$$U = \{U_1, U_2\}, U_1 \perp\!\!\!\perp U_2,$$

[15] Galles and Pearl (1997) proved recursive completeness assuming that, for any two variables, one knows which of the two (if any) is an ancestor of the other. Halpern (1998) proved recursive completeness without this assumption, provided only that (7.24) is known to hold for any two variables in the model. Halpern further provided a set of axioms for cases where the solution of $Y_x(u)$ is not unique or does not exist.

Task 3

Compute $P(Y_x = y)$ (i.e., the causal effect of smoking on cancer).

For any variable Z, by composition we have

$$Y_x(u) = Y_{xz}(u) \quad \text{if} \quad Z_x(u) = z.$$

Since $Y_{xz}(u) = Y_z(u)$ (from (7.29)),

Y_x (lowercase x)

$$Y_x(u) = Y_{xz_x}(u) = Y_z(u), \quad \text{where } z_x = Z_x(u). \tag{7.35}$$

Thus,

$$\begin{aligned} P(Y_x = y) &= P(Y_{zx} = y) && \text{from (7.35)} \\ &= \sum_z P(Y_{zx} = y \mid Z_x = z)\, P(Z_x = z) \\ &= \sum_z P(Y_z = y \mid Z_x = z)\, P(Z_x = z) && \text{by composition} \\ &= \sum_z P(Y_z = y)\, P(Z_x = z). && \text{from (7.30)} \end{aligned} \tag{7.36}$$

The probabilities $P(Y_z = y)$ and $P(Z_x = z)$ were computed in (7.34) and (7.31), respectively. Substituting gives us

$$P(Y_x = y) = \sum_z P(z \mid x) \sum_{x'} P(y \mid z, x')\ P(x'). \tag{7.37}$$

The right-hand side of (7.37) can be computed from $P(x, y, z)$ and coincides with the front-door formula derived in Section 3.4.3 (equation (3.42)).

Thus, $P(Y_x = y)$ can be reduced to expressions involving probabilities of observed variables and is therefore identifiable. More generally, our completeness result (Theorem 7.3.5) implies that *any* identifiable counterfactual quantity can be reduced to the correct expression by repeated application of composition and effectiveness (assuming recursiveness).

7.3.3 Axioms of Causal Relevance

In Section 1.2 we presented a set of axioms for a class of relations called *graphoids* (Pearl and Paz 1987; Geiger et al. 1990) that characterize informational relevance.[16] We now develop a parallel set of axioms for *causal relevance*, that is, the tendency of certain events to affect the occurrence of other events in the physical world, independent of the observer–reasoner. Informational relevance is concerned with questions of the form: "Given that we know Z, would gaining information about X give us new information

[16] "Relevance" will be used primarily as a generic name for the relationship of being relevant or irrelevant. It will be clear from the context when "relevance" is intended to negate "irrelevance."

seen that the meaning of the error term u_Y in the equation $Y = f_Y(pa_Y, u_Y)$ is captured by the counterfactual variable Y_{pa_Y}. In other words, the variable U_Y can be interpreted as a modifier of the functional mapping from PA_Y to Y. The statistics of such modifications is observable when pa_Y is held fixed. This translation into counterfactual notation may facilitate algebraic manipulations of U_Y without committing to the functional form of f_Y. However, from the viewpoint of model specification, the error terms should still be viewed as (summaries of) omitted factors.

Armed with this interpretation, we can obtain graphical and counterfactual definitions of causal concepts that were originally given error-based definitions. Examples of such concepts are causal influence, exogeneity, and instrumental variables (Section 5.4.3). In clarifying the relationships among error-based, counterfactual, and graphical definitions of these concepts, we should first note that these three modes of description can be organized in a simple hierarchy. Since graph separation implies independence, but independence does not imply graph separation (Theorem 1.2.4), definitions based on graph separation should imply those based on error-term independence. Likewise, since for any two variables X and Y the independence relation $U_X \perp\!\!\!\perp U_y$ implies the counterfactual independence $X_{pa_X} \perp\!\!\!\perp Y_{pa_Y}$ (but not the other way around), it follows that definitions based on error independence should imply those based on counterfactual independence. Overall, we have the following hierarchy:

$$\text{graphical criteria} \implies \text{error-based criteria} \implies \text{counterfactual criteria}.$$

The concept of exogeneity may serve to illustrate this hierarchy. The pragmatic definition of exogeneity is best formulated in counterfactual or interventional terms as follows.

Exogeneity (~~Counterfactual~~ Empirical Criterion)

A variable X is exogenous relative to Y if and only if the effect of X on Y is identical to the conditional probability of Y given X – that is, if

$$P(Y_x = y) = P(y \mid x) \tag{7.45}$$

or, equivalently,

$$P(Y = y \mid do\,(x)) = P(y \mid x); \tag{7.46}$$

this in turn ~~is equivalent to~~ follows from the independence condition $Y_x \perp\!\!\!\perp X$, named "weak ignorability" in Rosenbaum and Rubin (1983),[26] , though the converse may not hold.

This definition is pragmatic in that it highlights the reasons economists should be concerned with exogeneity by explicating the policy-analytic benefits of discovering that a variable is exogenous. However, this definition fails to guide an investigator toward

[26] We focus the discussion in this section on the causal component of exogeneity, which the econometric literature has unfortunately renamed "superexogeneity" (see Section 5.4.3). Epidemiologists refer to (7.46) as "no-confounding" (see (6.10)). We also postpone discussion of "strong ignorability," defined as the joint independence $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, to Chapter 9 (Definition 9.2.3).

verifying, from substantive knowledge of the domain, whether ~~this independence~~ the condition holds in any given system, especially when many equations are involved (see Section 11.3.2). To facilitate such judgments, economists (e.g., Koopmans 1950; Orcutt 1952) have adopted the error-based criterion of Definition 5.4.6.

Exogeneity (Error-Based Criterion)
A variable X is exogenous in M relative to Y if X is independent of all error terms that have an influence on Y ~~that is not mediated by X~~[27] when X is held constant.

This definition is more transparent to human judgment because the reference to error terms tends to focus attention on specific factors, potentially affecting Y, with which scientists are familiar. Still, to judge whether such factors are statistically independent is a difficult mental task unless the independencies considered are dictated by topological considerations that assure their stability. Indeed, the most popular conception of exogeneity is encapsulated in the notion of "common cause"; this may be stated formally as follows.

Exogeneity (Graphical Criterion)
A variable X is exogenous relative to Y if X and Y have no common ancestor in $G(M)$ or, equivalently, if all back-door paths between X and Y are blocked (by colliding arrows).[28]

It is easy to show that the graphical condition implies the error-based condition, which in turn implies the ~~counterfactual (or pragmatic)~~ empirical condition of (7.46). The converse implications do not hold. For example, Figure 6.4 illustrates a case where the graphical criterion fails and both the error-based and ~~counterfactual~~ empirical criteria classify X as exogenous. We argued in Section 6.4 that this type of exogeneity (there called "no confounding") is unstable or incidental, and we have raised the question of whether such cases were meant to be embraced by the definition. If we exclude unstable cases from consideration, then our three-level hierarchy collapses and all three definitions coincide.

(7.45)

Instrumental Variables: Three Definitions

A three-level hierarchy similarly characterizes the notion of instrumental variables (Bowden and Turkington 1984; Pearl 1995c; Angrist et al. 1996), illustrated in Figure 5.9. The traditional definition qualifies a variable Z as *instrumental* (relative to the pair (X, Y)) if (i) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X and (ii) Z is not independent of X.

---

[27] Independence relative to *all* errors is sometimes required in the literature (e.g., Dhrymes 1970, p. 169), but this is obviously too strong.

[28] As in Chapter 6 (note 19), the expression "common ancestors" should exclude nodes that have no other connection to Y except through X and should include latent nodes for every pair of dependent errors. Generalization to conditional exogeneity relative to observed covariates is straightforward in all three definitions.

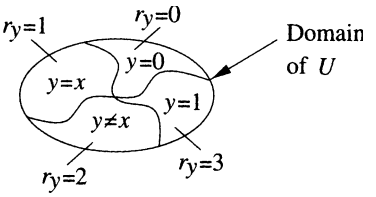

Figure 8.2 The canonical partition of U into four equivalence classes, each inducing a distinct functional mapping from X to Y for any given function $y = f(x, u)$.

Consider the structural equation that connects two binary variables, Y and X, in a causal model:

$$y = f(x, u).$$

For any given u, the relationship between X and Y must be one of four functions:

$$f_0 : y = 0, \quad f_1 : y = x,$$
$$f_2 : y \neq x, \quad f_3 : y = 1. \tag{8.5}$$

As u varies along its domain, regardless of how complex the variation, the only effect it can have on the model is to switch the relationship between X and Y among these four functions. This partitions the domain of U into four *equivalence classes*, as shown in Figure 8.2, where each class contains those points u that correspond to the same function. We can thus replace U by a four-state variable, $R(u)$, such that each state represents one of the four functions. The probability $P(u)$ would automatically translate into a probability function $P(r)$, $r = 0, 1, 2, 3$, that is given by the total weight assigned to the equivalence class corresponding to r. A state-minimal variable like R is called a "response" variable by Balke and Pearl (1994a,b) and a "mapping" variable by Heckerman and Shachter (1995), yet "canonical partition" would be more descriptive.[3]

Pearl (1993d) and

Because Z, X, and Y are all binary variables, the state space of U divides into 16 equivalence classes: each class dictates two functional mappings, one from Z to X and the other from X to Y. To describe these equivalence classes, it is convenient to regard each of them as a point in the joint space of two four-valued variables R_x and R_y. The variable R_x determines the compliance behavior of a subject through the mapping.

[3] In an experimental framework, this partition goes back to Greenland and Robins (1986) and was dubbed "Principal Stratification" by Frangakis and Rubin (2002). In this framework (see Section 7.4.4), u stands for an experimental unit and $R(u)$ corresponds to the potential response of unit u to treatment x. The assumption that each unit (e.g., an individual subject) possesses an intrinsic, seemingly "fatalistic" response function has met with some objections (Dawid 2000), owing to the inherent unobservability of the many factors that might govern an individual response to treatment. The equivalence-class formulation of $R(u)$ mitigates those objections (Pearl 2000) by showing that $R(u)$ evolves naturally and mathematically from any complex system of stochastic latent variables, provided only that we acknowledge their existence through the equation $y = f(x, u)$. Those who invoke quantum-mechanical objections to the latter step as well (e.g., Salmon 1998) should regard the functional relationship $y = f(x, u)$ as an abstract mathematical construct, representing the extreme points (vertices) of the set of conditional probabilities $P(y \mid x, u)$ satisfying the constraints of (8.1) and (8.2).

, 2011

Applying (8.6) and (8.7), we can write the linear transformation from a point $\vec{q}$ in Q to a point $\vec{p}$ in P:

$$\begin{aligned}
p_{00.0} &= q_{00} + q_{01} + q_{10} + q_{11}, \quad p_{00.1} = q_{00} + q_{01} + q_{20} + q_{21},\\
p_{01.0} &= q_{20} + q_{22} + q_{30} + q_{32}, \quad p_{01.1} = q_{10} + q_{12} + q_{30} + q_{32},\\
p_{10.0} &= q_{02} + q_{03} + q_{12} + q_{13}, \quad p_{10.1} = q_{02} + q_{03} + q_{22} + q_{23},\\
p_{11.0} &= q_{21} + q_{23} + q_{31} + q_{33}, \quad p_{11.1} = q_{11} + q_{13} + q_{31} + q_{33},
\end{aligned}$$

which can also be written in matrix form as $\vec{p} = \boldsymbol{R}\vec{q}$.

Given a point $\vec{p}$ in P-space, the strict lower bound on ACE($X \to Y$) can be determined by solving the following linear programming problem.

$$\text{Minimize } q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32}$$

subject to:

$$\begin{aligned}
&\sum_{j=0}^{3}\sum_{k=0}^{3} q_{jk} = 1,\\
&\boldsymbol{R}\vec{q} = \vec{p},\\
&q_{jk} \geq 0 \quad \text{for} \quad j, k \in \{0, 1, 2, 3\}.
\end{aligned} \tag{8.13}$$

For problems of this size, procedures are available for deriving symbolic expressions for the solution of this optimization exercise (Balke 1995), leading to the following lower bound on the treatment effect:

$$\text{ACE}(X \to Y) \geq \max \left\{ \begin{array}{c}
p_{11.1} + p_{00.0} - 1\\
p_{11.0} + p_{00.1} - 1\\
p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0}\\
p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1}\\
-p_{01.1} - p_{10.1}\\
-p_{01.0} - p_{10.0}\\
p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0}\\
p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1}
\end{array} \right\}. \tag{8.14a}$$

Similarly, the upper bound is given by

$$\text{ACE}(X \to Y) \leq \min \left\{ \begin{array}{c}
1 - p_{01.1} - p_{10.0}\\
1 - p_{01.0} - p_{10.1}\\
-p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0}\\
-p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0}\\
p_{11.1} + p_{00.1}\\
p_{11.0} + p_{00.0}\\
-p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0}\\
-p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1}
\end{array} \right\}. \tag{8.14b}$$

(replace minus with plus)

We may also derive bounds for (8.8) and (8.9) individually (under the same linear constraints), giving:

$$
\begin{aligned}
P(y_1 \mid do(x_0)) &\geq \max \begin{Bmatrix} p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{10.1} \\ p_{10.0} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{Bmatrix}, \\
P(y_1 \mid do(x_0)) &\leq \min \begin{Bmatrix} p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{Bmatrix},
\end{aligned} \tag{8.15}
$$

$$
\begin{aligned}
P(y_1 \mid do(x_1)) &\geq \max \begin{Bmatrix} p_{11.0} \\ p_{11.1} \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1} \\ -p_{01.0} - p_{10.0} + p_{10.1} + p_{11.1} \end{Bmatrix}, \\
P(y_1 \mid do(x_1)) &\leq \min \begin{Bmatrix} 1 - p_{01.0} \\ 1 - p_{01.1} \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} \end{Bmatrix}.
\end{aligned} \tag{8.16}
$$

1 - $p_{01.1}$

These expressions give the tightest possible assumption-free[4] bounds on the quantities sought.

8.2.4 The Natural Bounds

The expression for ACE($X \rightarrow Y$) (equation (8.4)) can be bounded by two simple formulas, each made up of the first two terms in (8.14a) and (8.14b) (Robins 1989; Manski 1990; Pearl 1994a):

$$
\begin{aligned}
\text{ACE}(X \rightarrow Y) &\geq P(y_1 \mid z_1) - P(y_1 \mid z_0) - P(y_1, x_0 \mid z_1) - P(y_0, x_1 \mid z_0), \\
\text{ACE}(X \rightarrow Y) &\leq P(y_1 \mid z_1) - P(y_1 \mid z_0) + P(y_0, x_0 \mid z_1) + P(y_1, x_1 \mid z_0).
\end{aligned} \tag{8.17}
$$

Because of their simplicity and wide range of applicability, the bounds given by (8.17) were named the *natural* bounds (Balke and Pearl 1997). The natural bounds guarantee that the causal effect of the actual treatment cannot be smaller than that of the encouragement ($P(y_1 \mid z_1) - P(y_1 \mid z_0)$) by more than the sum of two measurable quantities, $P(y_1, x_0 \mid z_1) + P(y_0, x_1 \mid z_0)$; they also guarantee that the causal effect of the treatment cannot exceed that of the encouragement by more than the sum of two other measurable

[4] "Assumption-transparent" might be a better term; we make no assumptions about factors that determine subjects' compliance, but we rely on the assumptions of (i) randomized assignment and (ii) no side effects, as displayed in the graph (e.g., Figure 8.1).

(lowercase italic "z")

quantities, $P(y_0, x_0 \mid z_1) + P(y_1, x_1 \mid z_0)$. The width of the natural bounds, not surprisingly, is given by the rate of noncompliance: $P(x_1 \mid z_0) + P(x_0 \mid z_1)$.

The width of the sharp bounds in (8.14ab) can be substantially narrower, though. In Balke (1995) and Pearl (1995b), it is shown that – even under conditions of 50% noncompliance – these bounds may collapse to a point and thus permit consistent estimation of ACE($X \rightarrow Y$). This occurs whenever (a) the percentage of subjects complying with assignment z_0 is the same as those complying with z_1 and (b) Y and Z are perfectly correlated in at least one treatment arm x (see Table 8.1 in Section 8.5).

Although more complicated than the natural bounds of (8.17), the sharp bounds of (8.14ab) are nevertheless easy to assess once we have the frequency data in the eight cells of $P(y, x \mid z)$. It can also be shown (Balke 1995) that the natural bounds are optimal when we can safely assume that no subject is *contrarian* – in other words, that no subject would consistently choose a treatment arm contrary to the one assigned.

Note that, if the response Y is continuous, then one can associate y_1 and y_0 with the binary events $Y > t$ and $Y \leq t$ (respectively) and let t vary continuously over the range of Y. Equations (8.15) and (8.16) would then provide bounds on the entire distribution of the treatment effect $P(Y < t \mid do(x))$.

8.2.5 Effect of Treatment on the Treated (ETT)

Much of the literature assumes that ACE($X \rightarrow Y$) is the parameter of interest, because ACE($X \rightarrow Y$) predicts the impact of applying the treatment uniformly (or randomly) over the population. However, if a policy maker is not interested in introducing new treatment policies but rather in deciding whether to maintain or terminate an existing program under its current incentive system, then the parameter of interest should measure the impact of the treatment *on the treated*, namely, the mean response of the treated subjects compared to the mean response of these same subjects had they not been treated (Heckman 1992). The appropriate formula for this parameter is

$$\begin{aligned} \text{ETT}(X \rightarrow Y) &= P(Y_{x_1} = y_1 \mid x_1) - P(Y_{x_0} = y_1 \mid x_1) \\ &= \sum_u [P(y_1 \mid x_1, u) - P(y_1 \mid x_0, u)]\, P(u \mid x_1), \end{aligned} \tag{8.18}$$

which is similar to (8.4) except for replacing the expectation over u with the conditional expectation given $X = x_1$.

The analysis of ETT($X \rightarrow Y$) reveals that, under conditions of *no intrusion* (i.e., $P(x_1 \mid z_0) = 0$, as in most clinical trials), ETT($X \rightarrow Y$) can be identified precisely (Bloom 1984; Heckman and Robb 1986; Angrist and Imbens 1991). The natural bounds governing ETT($X \rightarrow Y$) in the general case can be obtained by similar means (Pearl 1995b), which yield

$$\begin{aligned} \text{ETT}(X \rightarrow Y) &\geq \frac{P(y_1 \mid z_1) - P(y_1 \mid z_0)}{P(x_1)/P(z_1)} - \frac{P(y_0, x_1 \mid z_0)}{P(x_1)}, \\ \text{ETT}(X \rightarrow Y) &\leq \frac{P(y_1 \mid z_1) - P(y_1 \mid z_0)}{P(x_1)/P(z_1)} + \frac{P(y_1, x_1 \mid z_0)}{P(x_1)}. \end{aligned} \tag{8.19}$$

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

$$P(y_1 \mid x_1) - p(y_1 \mid x_0) = \frac{0.473}{0.473 + 0.139} - \frac{0.073 + 0.081}{1 + 0.315 + 0.073} = 0.662,$$

and an encouragement effect (intent to treat) of

$$P(y_1 \mid z_1) - P(y_1 \mid z_0) = 0.073 + 0.473 - 0.081 = 0.465.$$

According to (8.17), ACE($X \rightarrow Y$) can be bounded by

$$\text{ACE} \geq \text{ETT}(X \rightarrow Y) \geq 0.465 - 0.073 - 0.000 = 0.392,$$
$$\text{ACE} \leq \text{ETT}(X \rightarrow Y) \leq 0.465 + 0.315 + 0.000 = 0.780.$$

These are remarkably informative bounds: although 38.8% of the subjects deviated from their treatment protocol, the experimenter can categorically state that, when applied uniformly to the population, the treatment is guaranteed to increase by at least 39.2% the probability of reducing the level of cholesterol by 28 points or more.

The impact of treatment "on the treated" is equally revealing. Using equation (8.20), ETT($X \rightarrow Y$) can be evaluated precisely (since $P(x_1 \mid z_0) = 0$):

$$\text{ETT}(X \rightarrow Y) = \frac{0.465}{0.610} = 0.762.$$

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

8.3 COUNTERFACTUALS AND LEGAL RESPONSIBILITY

Evaluation of counterfactual probabilities could be enlightening in some legal cases in which a plaintiff claims that a defendant's actions were responsible for the plaintiff's misfortune. Improper rulings can easily be issued without an adequate treatment of counterfactuals (Robins and Greenland 1989). Consider the following hypothetical and fictitious case study, specially crafted in Balke and Pearl (1994a) to accentuate the disparity between causal effects and causal attribution.

The marketer of PeptAid (antacid medication) randomly mailed out product samples to 10% of the households in the city of Stress, California. In a follow-up study, researchers determined for each individual whether they received the PeptAid sample, whether they consumed PeptAid, and whether they developed peptic ulcers in the following month.

The causal structure for this scenario is identical to the partial compliance model given by Figure 8.1, where z_1 asserts that PeptAid was received from the marketer, x_1 asserts that PeptAid was consumed, and y_1 asserts that peptic ulceration occurred. The data showed the following distribution:

$$P(y_0, x_0 \mid z_0) = 0.32, \quad P(y_0, x_0 \mid z_1) = 0.02,$$

$$P(y_0, x_1 \mid z_0) = 0.32, \quad P(y_0, x_1 \mid z_1) = 0.17,$$

$$P(y_1, x_0 \mid z_0) = 0.04, \quad P(y_1, x_0 \mid z_1) = 0.67,$$

$$P(y_1, x_1 \mid z_0) = 0.32, \quad P(y_1, x_1 \mid z_1) = 0.14.$$

These data indicate a high correlation between those who consumed PeptAid and those who developed peptic ulcers:

$$P(y_1 \mid x_1) = 0.50, \quad P(y_1 \mid x_0) = 0.26.$$

In addition, the intent-to-treat analysis showed that those individuals who received the PeptAid samples had a 45% greater chance of developing peptic ulcers:

$$P(y_1 \mid z_1) = 0.81, \quad P(y_1 \mid z_0) = 0.36.$$

The plaintiff (Mr. Smith), having heard of the study, litigated against both the marketing firm and the PeptAid producer. The plaintiff's attorney argued against the producer, claiming that the consumption of PeptAid triggered his client's ulcer and resulting medical expenses. Likewise, the plaintiff's attorney argued against the marketer, claiming that his client would not have developed an ulcer if the marketer had not distributed the product samples.

The defense attorney, representing both the manufacturer and marketer of PeptAid, rebutted this argument, stating that the high correlation between PeptAid consumption and ulcers was attributable to a common factor, namely, pre-ulcer discomfort. Individuals with gastrointestinal discomfort would be much more likely both to use PeptAid and to develop stomach ulcers. To bolster his clients' claims, the defense attorney introduced expert analysis of the data showing that, on average, consumption of PeptAid actually decreases an individual's chances of developing ulcers by at least 15%.

Indeed, the application of (8.14a,b) results in the following bounds on the average causal effect of PeptAid consumption on peptic ulceration:

(replace "ETT" with "ACE")

$$-0.23 \leq \text{~~ETT~~ ACE}(X \rightarrow Y) \leq -0.15;$$

this proves that PeptAid is beneficial to the population as a whole.

The plaintiff's attorney, though, stressed the distinction between the average treatment effects for the entire population and for the subpopulation consisting of those individuals who, like his client, received the PeptAid sample, consumed it, and then developed ulcers. Analysis of the population data indicated that, had PeptAid not been distributed, Mr. Smith would have had at most a 7% chance of developing ulcers – regardless of any confounding factors such as pre-ulcer pain. Likewise, if Mr. Smith had not consumed PeptAid, he would have had at most a 7% chance of developing ulcers.

The damaging statistics against the marketer are obtained by evaluating the bounds on the counterfactual probability that the plaintiff would have developed a peptic ulcer if he had not received the PeptAid sample, given that he in fact received the sample PeptAid, consumed the PeptAid, and developed peptic ulcers. This probability may be written in terms of the parameters q_{13}, q_{31}, and q_{33} as

Table 9.1

	Exposure	
	High (x)	Low (x')
Deaths (y)	30	16
Survivals (y')	69,130	59,010

matching our intuition that a shot fired by an expert marksman would be sufficient for causing the death of T, regardless of the court decision.

Note that Theorems 9.2.10 and 9.2.11 are not applicable to this example because x is not exogenous; events x and y have a common cause (the captain's signal), which renders $P(y \mid x') = 0 \neq P(y_{x'}) = \frac{1}{2}$. However, the monotonicity of Y (in x) permits us to compute PNS, PS, and PN from the joint distribution $P(x, y)$ and the causal effects (using (9.28)–(9.30)), instead of consulting the functional model. Indeed, writing

$$P(x, y) = P(x', y') = \tfrac{1}{2} \tag{9.40}$$

and

$$P(x, y') = P(x', y) = 0, \tag{9.41}$$

we obtain

$$\text{PN} = \frac{P(y) - P(y_{x'})}{P(x, y)} = \frac{\frac{1}{2} - \frac{1}{2}}{\frac{1}{2}} = 0 \tag{9.42}$$

and

$$\text{PS} = \frac{P(y_x) - P(y)}{P(x', y')} = \frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1, \tag{9.43}$$

as expected.

9.3.3 Example 3: ~~The Effect of Radiation on Leukemia~~ Was Radiation the Cause of Leukemia?

Consider the following data (Table 9.1, adapted[10] from Finkelstein and Levin 1990) comparing leukemia deaths in children in southern Utah with high and low exposure to radiation from the fallout of nuclear tests in Nevada. Given these data, we wish to estimate the probabilities that high exposure to radiation was a necessary (or sufficient, or both) cause of death due to leukemia.

[10] The data in Finkelstein and Levin (1990) are given in "person-year" units. For the purpose of illustration we have converted the data to absolute numbers (of deaths and nondeaths) assuming a ten-year observation period.

Table 9.3. ***PN as a Function of Assumptions and Available Data***

Assumptions			Data Available		
Exogeneity	Monotonicity	Additional	Experimental	Observational	Combined
+	+		ERR	ERR	ERR
+	−		bounds	bounds	bounds
−	+	covariate control	—	corrected ERR	corrected ERR
−	+		—	—	corrected ERR
−	−		—	—	bounds

Note: ERR stands for the excess risk ratio, $1 - P(y \mid x)/P(y' \mid x')$; corrected ERR is given in (9.31).

Note: this "/" was erroneously changed to ">" in the last printing - it should be changed backed to "/"

can be ascertained: exogeneity (i.e., no confounding) and monotonicity (i.e., no prevention). When monotonicity does not hold, ERR provides merely a lower bound for PN, as shown in (9.13). (The upper bound is usually unity.) The nonentries (—) in the right-hand side of Table 9.3 represent vacuous bounds (i.e., $0 \leq \text{PN} \leq 1$). In the presence of confounding, ERR must be corrected by the additive term $[P(y \mid x') - P(y_{x'})]/P(x, y)$, as stated in (9.31). In other words, when confounding bias (of the causal effect) is positive, PN is higher than ERR by the amount of this additive term. Clearly, owing to the division by $P(x, y)$, the PN bias can be many times higher than the causal effect bias $P(y \mid x') - P(y_{x'})$. However, confounding results only from association between exposure and other factors that affect the outcome; one need not be concerned with associations between such factors and susceptibility to exposure (see Figure 9.2).

The last row in Table 9.3, corresponding to no assumptions whatsoever, leads to vacuous bounds for PN, unless we have combined data. This does not mean, however, that justifiable assumptions *other* than monotonicity and exogeneity could not be helpful in rendering PN identifiable. The use of such assumptions is explored in the next section.

9.4 IDENTIFICATION IN NONMONOTONIC MODELS

In this section we discuss the identification of probabilities of causation without making the assumption of monotonicity. We will assume that we are given a causal model M in which all functional relationships are known, but since the background variables U are not observed, their distribution is not known and the model specification is not complete.

Our first step would be to study under what conditions the function $P(u)$ can be identified, thus rendering the entire model identifiable. If M is Markovian, then the problem can be analyzed by considering each parents–child family separately. Consider any arbitrary equation in M,

$$\begin{aligned} y &= f(pa_Y, u_Y) \\ &= f(x_1, x_2, \ldots, x_k, u_1, \ldots, u_m), \end{aligned} \tag{9.55}$$

made assumption of "no prevention" and for the often asked question of whether a clinical study is representative of its target population (equation (9.32)).

On the conceptual side, we have seen that both the probability of necessity (PN) and probability of sufficiency (PS) play a role in our understanding of causation and that each component has its logic and computational rules. Although the counterfactual concept of necessary cause (i.e., that an outcome would not have occurred "but for" the action) is predominant in legal settings (Robertson 1997) and in ordinary discourse, the sufficiency component of causation has a definite influence on causal thoughts.

The importance of the sufficiency component can be uncovered in examples where the necessary component is either dormant or ensured. Why do we consider striking a match to be a more adequate explanation (of a fire) than the presence of oxygen? Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact, the PN is actually higher for the oxygen if we allow for alternative ways of igniting a spark.) Thus, it must be the sufficiency component that endows the match with greater explanatory power than the oxygen. If the probabilities associated with striking a match and the presence of oxygen are denoted p_m and p_o, respectively, then the PS measures associated with these explanations evaluate to PS(match) $\approx p_o$ and PS(oxygen) $\approx p_m$, clearly favoring the match when $p_o \gg P_m$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

Should PS enter legal considerations in criminal and tort law? I believe that it should – as does Good (1993) – because attention to sufficiency implies attention to the consequences of one's action. The person who lighted the match ought to have anticipated the presence of oxygen, whereas the person who supplied – or could (but did not) remove – the oxygen is not generally expected to have anticipated match-striking ceremonies.

However, what weight should the law assign to the necessary versus the sufficient component of causation? This question obviously lies beyond the scope of our investigation, and it is not at all clear who would be qualified to tackle the issue or whether our legal system would be prepared to implement the recommendation. I am hopeful, however, that whoever undertakes to consider such questions will find the analysis in this chapter to be of some use. The next chapter combines aspects of necessity and sufficiency in explicating a more refined notion: "actual cause."

Acknowledgments

I am indebted to Sander Greenland for many suggestions and discussions concerning the treatment of attribution in the epidemiological literature and the potential applications of our results in practical epidemiological studies. Donald Michie and Jack Good are responsible for shifting my attention from PN to PS and PNS. Clark Glymour and Patricia Cheng helped to unravel some of the mysteries of causal power theory, and Michelle Pearl provided useful pointers to the epidemiological literature. Blai Bonet corrected omissions from earlier versions of Lemmas 9.2.7 and 9.2.8, and Jin Tian tied it all up in tight bounds.

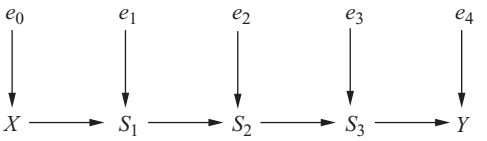

Figure 11.4 Showing the noise factors on the path from X to Y.

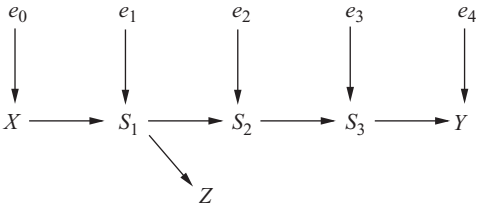

Figure 11.5 Conditioning on Z creates dependence between X and e_1, which biases the estimated effect of X on Y.

Author's Answer:

The exclusion of descendants from the back-door criterion is indeed based on first principles, in terms of the goal of removing bias. The principles are as follows: We wish to measure a certain quantity (causal effect) and, instead, we measure a dependency $P(y \mid x)$ that results from all the paths in the diagram; some are spurious (the back-door paths), and some are genuinely causal (the directed paths from X to Y). Thus, to remove bias, we need to modify the measured dependency and make it equal to the desired quantity. To do this systematically, we condition on a set Z of variables while ensuring that:

1. We block all spurious paths from X to Y,
2. We leave all directed paths unperturbed,
3. We create no new spurious paths.

Principles 1 and 2 are accomplished by blocking all back-door paths and only those paths, as articulated in condition (ii). Principle 3 requires that we do not condition on descendants of X, even those that do not block directed paths, because such descendants may create new spurious paths between X and Y. To see why, consider the graph

$$X \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow Y.$$

The intermediate variables, S_1, S_2,..., (as well as Y) are affected by noise factors e_0, e_1, e_2,. . . which are not shown explicitly in the diagram. However, under magnification, the chain unfolds into the graph in Figure 11.4.

Now imagine that we condition on a descendant Z of S_1 as shown in Figure 11.5. Since S_1 is a ~~collider~~, this creates dependency between X and e_1 which is equivalent to a back-door path

italics

virtual collider

$$X \leftrightarrow e_1 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow Y.$$

By principle 3, such paths should not be created, for it introduces spurious dependence between X and Y.

Note that a descendant Z of X that is not also a descendant of some S_i escapes this exclusion; it can safely be conditioned on without introducing bias (though it may decrease the efficiency of the associated estimator of the causal effect of X on Y). Section

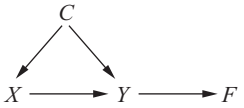

Figure 11.6 Graph applicable for accessing the effect of X on Y.

because the graph applicable for this task is given in Figure 11.6; F becomes a descendant of X, and is excluded by the back-door criterion.

2. If the explanation of confounding and sufficiency sounds at variance with traditional epidemiology, it is only because traditional epidemiologists did not have proper means of expressing the operations of blocking or creating dependencies. They might have had a healthy intuition about dependencies, but graphs translate this intuition into a formal system of closing and opening paths.

   We should also note that before 1985, causal analysis in epidemiology was in a state of confusion, because the healthy intuitions of leading epidemiologists had to be expressed in the language of associations – an impossible task. Even the idea that confounding stands for "bias," namely, a "difference between two dependencies, one that we wish to measure, the other that we do measure," was resisted by many (see Chapter 6), because they could not express the former mathematically.[3]

   Therefore, instead of finding "something in graph language that is closer to traditional meaning," we can do better: explicate what that "traditional meaning" ought to have been.

   In other words, traditional meaning was informal and occasionally misguided, while graphical criteria are formal and mathematically proven.

   Chapter 6 (pp. 183, 194) records a long history of epidemiological intuitions, some by prominent epidemiologists, that have gone astray when confronted with questions of confounding and adjustment (see Greenland and Robins 1986; Wickramaratne and Holford 1987; Weinberg 1993). Although most leading epidemiologists today are keenly attuned to modern developments in causal analysis, (e.g., Glymour and Greenland 2008), epidemiological folklore is still permeated with traditional intuitions that are highly suspect. (See Section 6.5.2.)

   In summary, graphical criteria, as well as principles 1–3 above, give us a sensible, friendly, and unambiguous interpretation of the "traditional meaning of epidemiological concepts."

11.3.2 Demystifying "Strong Ignorability"

Researchers working within the confines of the potential-outcome language express the condition of "zero bias" or "no-confounding" using an independence relationship called

[3] Recall that Greenland and Robins (1986) were a lone beacon of truth for many years, and even they had to resort to the "black-box" language of "exchangeability" to define "bias," which discouraged intuitive interpretations of confounding (see Section 6.5.3). Indeed, it took epidemiologists another six years (Weinberg 1993) to ~~discover that adjusting for factors affected by the exposure~~ (as in Figure 11.5) ~~would introduce bias.~~

notice that bias could be induced by adjusting for a factor affected by a mediator of the effect under study

(2010) provides a complete criterion for *c*-equivalence.

$S_2 = \{Z_2, W_2\}$ is admissible (by virtue of satisfying the back-door criterion), hence S_1 and S_2 are *c*-equivalent. Yet neither C_1 nor C_2 holds in this case.

A natural attempt would be to impose the condition that S_1 and S_2 each be *c*-equivalent to $S_1 \cup S_2$ and invoke the criterion of Stone (1993) and Robins (1997) for the required set-subset equivalence. The resulting criterion, while valid, is still not complete; there are cases where S_1 and S_2 are *c*-equivalent yet not *c*-equivalent to their union. A theorem by Pearl and Paz ~~(2008) broadens this condition using irreducible sets.~~

Having given a conditional-independence characterization of *c*-equivalence does not solve, of course, the problem of identifying admissible sets; the latter is a causal notion and cannot be given statistical characterization.

The graph depicted in Figure 11.8(b) demonstrates the difficulties commonly faced by social and health scientists. Suppose our target is to estimate $P(y \mid do(x))$ given measurements on $\{X, Y, Z_1, Z_2, W_1, W_2, V\}$, but having no idea of the underlying graph structure. The conventional wisdom is to start with all available covariates $C = \{Z_1, Z_2, W_1, W_2, V\}$, and test if a proper subset of C would yield an equivalent estimand upon adjustment. Statistical methods for such reduction are described in Greenland et al. (1999b), Geng et al. (2002), and Wang et al. (2008). For example, $\{Z_1, V\}$, $\{Z_2, V\}$, or $\{Z_1, Z_2\}$ can be removed from C by successively applying conditions C_1 and C_2. This reduction method would produce three irreducible subsets, $\{Z_1, W_1, W_2\}$, $\{Z_2, W_1, W_2\}$, and $\{V, W_1, W_2\}$, all *c*-equivalent to the original covariate set C. However, none of these subsets is admissible for adjustment, because none (including C) satisfies the back-door criterion. While a theorem due to Tian et al. (1998) assures us that any *c*-equivalent subset of a set C can be reached from C by a step-at-a-time removal method, going through a sequence of *c*-equivalent subsets, the problem of covariate selection is that, lacking the graph structure, we do not know which (if any) of the many subsets of C is admissible. The next subsection discusses how external knowledge, as well as more refined analysis of the data at hand, can be brought to bear on the problem.

11.3.4 Data vs. Knowledge in Covariate Selection

What then can be done in the absence of a causal graph? One way is to postulate a plausible graph, based on one's understanding of the domain, and check if the data refutes any of the statistical claims implied by that graph. In our case, the graph of Figure 11.8(b) advertises several such claims, cast as conditional independence constraints, each associated with a missing arrow in the graph:

$$
\begin{array}{lll}
V \perp\!\!\!\perp X \mid Z_1, W_1 & V \perp\!\!\!\perp W_2 & Z_2 \perp\!\!\!\perp W_1 \mid W_2 \\
V \perp\!\!\!\perp Y \mid X, Z_2, W_2 & Z_1 \perp\!\!\!\perp Z_2 \mid V, W_1, W_2 & Z_1 \perp\!\!\!\perp W_2 \mid W_1 \\
V \perp\!\!\!\perp W_1 & X \perp\!\!\!\perp Z_2 \mid Z, W_1, W_2 & X \perp\!\!\!\perp \{V, Z_2\} \mid Z_1, W_1, W_2.
\end{array}
$$

Satisfying these constraints does not establish, of course, the validity of the causal model postulated because, as we have seen in Chapter 2, alternative models may exist which satisfy the same independence constraints yet embody markedly different causal structures, hence, markedly different admissible sets and effect estimands. A trivial example would be a complete graph, with arbitrary orientation of arrows which, with a clever choice of parameters, can emulate any other graph. A less trivial example, one that is not sensitive to choice of parameters, lies in the class of equivalent structures, in

The difficulty that most investigators experience in comprehending what "ignorability" means, and what judgment it summons them to exercise, has tempted them to assume that it is automatically satisfied, or at least is likely to be satisfied, if one includes in the analysis as many covariates as possible. The prevailing attitude is that adding more covariates can cause no harm (Rosenbaum 2002, p. 76) and can absolve one from thinking about the causal relationships among those covariates, the treatment, the outcome and, most importantly, the confounders left unmeasured (Rubin 2009).

This attitude stands contrary to what students of graphical models have learned, and what this book has attempted to teach. The admissibility of S can be established only by appealing to the causal knowledge available to the investigator, and that knowledge, as we know from graph theory and the back-door criterion, makes bias reduction a nonmonotonic operation, i.e., eliminating bias (or imbalance) due to one confounder may awaken and unleash bias due to dormant, unmeasured confounders. Examples abound (e.g., Figure 6.3) where adding a variable to the analysis not only is not needed, but would introduce irreparable bias (Pearl 2009, Shrier 2009, Sjölander 2009[a]). 2009a

Another factor inflaming the controversy has been the general belief that the bias-reducing potential of propensity score methods can be assessed experimentally by running case studies and comparing effect estimates obtained by propensity scores to those obtained by controlled randomized experiments (Shadish and Cook 2009).[11] This belief is unjustified because the bias-reducing potential of propensity scores depends critically on the specific choice of S or, more accurately, on the cause–effect relationships among variables inside and outside S. Measuring significant bias in one problem instance (say, an educational program in Oklahoma) does not preclude finding zero bias in another (say, crime control in Arkansas), even under identical statistical distributions $P(x, s, y)$.

With these considerations in mind, one is justified in asking a social science type question: What is it about propensity scores that has inhibited a more general understanding of their promise and limitations?

Richard Berk, in *Regression Analysis: A Constructive Critique* (Berk 2004), recalls similar phenomena in social science, where immaculate ideas were misinterpreted by the scientific community: "I recall a conversation with Don Campbell in which he openly wished that he had never written Campbell and Stanley (1966). The intent of the justly famous book, *Experimental and Quasi-Experimental Designs for Research,* was to contrast randomized experiments to quasi-experimental approximations and to strongly discourage the latter. Yet the apparent impact of the book was to legitimize a host of quasi-experimental designs for a wide variety of applied social science. After I got to know Dudley Duncan late in his career, he said that he often thought that his influential book on path analysis, *Introduction to Structural Equation Models* was a big mistake. Researchers had come away from the book believing that fundamental policy questions about social inequality could be quickly and easily answered with path analysis." (p. xvii)

[11] Such beliefs are encouraged by valiant statements such as: "For dramatic evidence that such an analysis can reach the same conclusion as an exactly parallel randomized experiment, see Shadish and Clark (2006, unpublished)" (Rubin 2007).

I believe that a similar cultural phenomenon has evolved around propensity scores.

It is not that Rosenbaum and Rubin were careless in stating the conditions for success. Formally, they were very clear in warning practitioners that propensity scores work only under "strong ignorability" conditions. However, what they failed to realize is that it is not enough to warn people against dangers they cannot recognize; to protect them from perilous adventures, we must also give them eyeglasses to spot the threats, and a meaningful language to reason about them. By failing to equip readers with tools (e.g., graphs) for recognizing how "strong ignorability" can be violated or achieved, they have encouraged a generation of researchers (including federal agencies) to assume that ignorability either holds in most cases, or can be made to hold by clever designs.

11.3.6 The Intuition behind *do*-Calculus

Question to Author Regarding Theorem 3.4.1:

In the inference rules of *do*-calculus (p. 85), the subgraph $G_{\overline{X}}$, represents the distribution prevailing under the operation $do(X = x)$, since all direct causes of X are removed. What distribution does the submodel $G_{\underline{X}}$ represent, with the direct effects of X removed?

Author's Reply:

The graph $G_{\underline{X}}$ represents the hypothetical act of "holding constant" all children of X. This severs all directed paths from X to Y, while leaving all back-door paths intact. So, if X and Y are d-connected in that graph, it must be due to (unblocked) confounding paths between the two. Conversely, if we find a set Z of nodes that d-separate X from Y in that graph, we are assured that Z blocks all back-door paths in the original graph. If we further condition on variables Z, we are assured, by the back-door criterion, that we have neutralized all confounders and that whatever dependence we measure after such conditioning must be due to the causal effect of X on Y, free of confoundings.

11.3.7 The Validity of *G*-Estimation

"*g*" in italics

g-formula

In Section 3.6.4 we introduced the ~~*G*-estimation formula~~ (3.63), together with the counterfactual independency (3.62), $(Y(x) \perp\!\!\!\perp X_k \mid \overline{L}_k, \overline{X}_{k-1} = \overline{x}_{k-1})$, which Robins proved to be a sufficient condition for (3.63). In general, condition (3.62) ~~is both~~ overrestrictive and ~~lacks~~ intuitive basis. A ~~more general and intuitive~~ condition leading to (3.63) is derived in (4.5) (p. 122), which reads as follows:

lacking

may be

less restrictive graphical

(3.62*) ~~General Condition for g-Estimation~~ (Sequential ~~Deconfounding~~) Back-door)

$P(y \mid g = x)$ is identifiable and is given by (3.63) *if every action-avoiding back-door path from X_k to Y is blocked by some subset L_k of nondescendants of X_k.* (By "action-avoiding" we mean a path containing no arrows entering an X variable later than X_k.)

X_k,...,X_n.

replace paragraph

~~Th~~ The two conditions are compared in the following examples. ~~onstrated in the fol-~~ ~~lowing three examples.~~

Example 11.3.1 Figure 11.10 demonstrates cases where the g-formula (3.63) is valid with a subset L_k of the past but not with the entire past. Assuming U_1 and U_2 are

, in some topological ordering of the variables.)

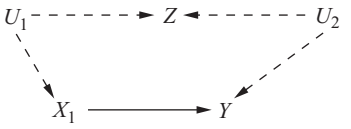

Figure 11.10 Conditioning on the entire past $L_1 = Z$ would invalidate g-estimation.

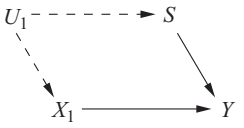

Figure 11.11 g-estimation is rendered valid by including a non-predecessor S.

unobserved, and temporal order: U_1, Z, X_1, U_2, Y, we see that both (3.62) and (3.62*), hence (3.63), are satisfied with $L_1 = 0$, while taking the whole past $L_1 = Z$ would violate both. (Robins (1986) did not insist on temporal ordering; it is implied in (Robins, 1995).)

Example 11.3.2 Figure 11.11 demonstrates cases where defining L_k as the set of "nondescendants" of X_k (as opposed to temporal predecessors of X_k) broadens (3.62). Assuming temporal order: U_1, X_1, S, Y, both (3.62) and (3.62*) are satisfied with $L_1 = S$, but not with $L_1 = 0$.

replace with corrected Example 11.3.3

Example 11.3.3 In a previous edition of this book (2009) it was argued that Figure 11.12 demonstrates cases where (3.62) is not satisfied, while the graphical condition of (3.62*) is. A recent analysis by Richardson and Robins (Working Paper Number 128, Center for Statistics and Social Sciences, University of Washington, April 2013) shows this not to be the case. Since condition (3.62) refers to one specific instantiation of $\overline{X}_{k-1}$, not to $\overline{X}_{k-1}$ as a variable, it is satisfied in the graph of Fig. 11.11. I am grateful to Richardson and Robins for this illuminating observation.

replace two sentences

A more serious weakness of articulating scientific assumptions in the language of counterfactuals (or potential outcomes) is opacity. The counterfactual condition (3.62) that legitimizes the use of the g-formula evokes no scientific context to judge the plausibility of the condition. ~~port.~~ Epidemiologists who apply this formula are doing so under no guidance of substantive medical knowledge. Fortunately, graphical methods are rapidly making their way into epidemiological practice (Greenland et al. 1999a; Robins 2001; Hernán et al. 2002; Greenland and Brumback 2002; Kaufman et al. 2005; Petersen et al. 2006; VanderWeele and Robins 2007) as more ~~and more~~ researchers begin to understand the assumptions behind ~~g-estimation. With the added understanding~~ that structural ~~equation~~ causal models subsume, unify, and underlie the graphical, counterfactual, and potential outcome ~~and sufficient component (Rothman 1976)~~ approaches to causation,[12] epidemiology stands a good chance of becoming the first discipline to fully liberate itself from past dogmas and "break through our academically enforced reluctance to think directly about causes (Weinberg 2007)."

their estimation routines. Given the general recognition

[12] Thi
Bru
Wir
resp
tura

replace footnote 12 text

Connections between SEMs and potential outcomes have been noted in many fields, including epidemiology (Greenland and Brumback 2002), economics (Heckman and Vytlacil 2007), sociology (Morgan and Winship 2007), and even statistics (Cox and Wermuth 2004). We here wish to emphasize further that the two are one and the same species (logically isomorphic), so that a theorem in one is a theorem in the other; as a consequence, SEMs provide the semantic basis as well a parsimonous and transparent representation for potential outcomes (Chapter 7).

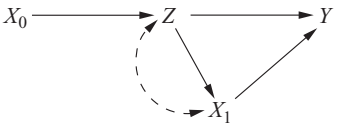

Figure 11.12 A graph for which ~~g-estimation is valid while Robins' condition (3.62) is violated.~~ g and (3.62*) are both valid.

11.4 POLICY EVALUATION AND THE *do*-OPERATOR

11.4.1 Identifying Conditional Plans (Section 4.2, p. 113)

Question to Author:

Section 4.2 of the book (p. 113) gives an identification condition and estimation formula for the effect of a conditional action, namely, the effect of an action $do(X = g(z))$ where $Z = z$ is a measurement taken prior to the action. Is this equation generalizable to the case of several actions, i.e., conditional plan?

The difficulty seen is that this formula was derived on the assumption that X does not change the value of Z. However, in a multiaction plan, some actions in X could change observations Z that guide future actions. We do not have notation for distinguishing post-intervention from pre-intervention observations. Absent such notation, it is not clear how conditional plans can be expressed formally and submitted to the *do*-calculus for analysis.

***Author's Reply* (with Ilya Shpitser):**

A notational distinction between post-intervention pre-intervention observations is introduced in Chapter 7 using the language of counterfactuals. The case of conditional plans, however, can be handled without resorting to richer notation. The reason is that the observations that dictate the choice of an action are not changed by that action, while those that have been changed by previous actions are well captured by the $P(y \mid do(x), z)$ notation.

To see that this is the case, however, we will first introduce counterfactual notation, and then show that it can be eliminated from our expression. We will use bold letters to denote sets, and normal letters to denote individual elements. Also, capital letters will denote random variables, and small letters will denote possible values these variables could attain. We will write Y_x to mean 'the value Y attains if we set variables X to values x.' Similarly, Y_{Xg} is taken to mean 'the value Y attains if we set variables X to whatever values they would have attained under the stochastic policy g.' Note that Y_x and Y_{Xg} are both random variables, just as the original variable Y.

Say we have a set of K action variables X that occur in some temporal order. We will indicate the time at which a given variable is acted on by a superscript, so a variable X^i occurs before X^j if $i < j$. For a given X^i, we denote $X^{<i}$ to be the set of action variables preceding X^i.

We are interested in the probability distribution of a set of outcome variables Y, under a policy that sets the values of each $X^i \in X$ to the output of functions g_i (known in advance) which pay attention to some set of prior variables Z_i, as well as the previous interventions on $X^{<i}$. At the same time, the variables Z^i are themselves affected by previous interventions. To define this recursion appropriately, we use an inductive definition. The base case is $X^1_g = g_1(Z_1)$. The inductive case is $X^i_g = g_i(Z^i_{X_g^{<i}}, X_g^{<i})$. Here the

one can hardly find a serious discussion of what the model means, once it is fitted and estimated (see Section 11.5.3 for SEM survival kit).[13]

The weakness of this educational tradition surfaces when inquisitive students ask questions that deviate slightly from standard LISREL routines, the answers to which hinge on the causal interpretation of structural coefficients and structural equations. For example:

1. Why should we define the total effect the way we do? (i.e., as the sum of products of certain direct effects). Is this an arbitrary definition, or is it compelled by the causal interpretation of the path coefficients?
2. Why should we define the indirect effect as the difference between the total and direct effects?
3. How can we define direct effect in nonlinear systems or in systems involving dichotomous variables?
4. How should we, in a meaningful way, define effects in systems involving feedback loops (i.e., reciprocal causation) so as to avoid the pitfalls of erroneous definitions quoted in SEM textbooks? (see p. 164)
5. Would our assessment of direct and total effects remain the same if we were to take some measurements prior to implementing the action whose effect we attempt to estimate?

Readers will be pleased to note that these questions can be given formal answers, as in Sections 4.5.4, 4.5.5, 11.5.2, 11.5.3, and 11.7.1.

On a personal note, my interest in direct and indirect effects was triggered by a message from Jacques Hagenaars, who wrote (September 15, 2000): "Indirect effects do occupy an important place in substantive theories. Many social science theories 'agree' on the input (background characteristics) and output (behavioral) variables, but differ exactly with regard to the intervening mechanisms. To take a simple example, we know that the influence of Education on Political Preferences is mediated through 'economic status' (higher educated people get the better jobs and earn more money) and through a 'cultural mechanism' (having to do with the contents of the education and the accompanying socialization processes at school). We need to know and separate the nature and consequences of these two different processes, that is, we want to know the signs and the magnitudes of the indirect effects. In the parametric linear version of structural equation models, there exists a 'calculus of path coefficients' in which we can write total effects in terms of direct and several indirect effects. But this is not possible in the general nonparametric cases and not, e.g., in the log-linear parametric version. For systems of logic models there does not exist a comparable 'calculus of path coefficients' as has been remarked long ago. However, given its overriding theoretical importance, the issue of indirect effects cannot be simply neglected."

Stimulated by these comments, and armed with the notation of nested counterfactuals, I set out to formalize the legal definition of hiring discrimination given on page ~~147~~ 131, and

[13] The word "causal" does not appear in the index of any of the post-2000 SEM textbooks that I have examined.

It should not be too hard to convince our Bayesian that these two assessments could not be totally arbitrary, but must obey some restrictions of coherence. For example, the inequality $P(y \mid do(x)) \geq P(y, x)$ should be obeyed for all events x and y.[17] Moreover, coherence restrictions of this kind are automatically satisfied whenever $P(y \mid do(x))$ is derived from a causal network according to the rules of Chapter 3. These two arguments should be inviting for a Bayesian to start drawing mathematical benefits from causal calculus, while maintaining caution and skepticism, and, as they say in the Talmud:

> "*From benefits comes understanding*"
> (free translation of "*mitoch shelo lishma, ba lishma*" (Talmud, Psahim, 50b)).

Bayesians will eventually embrace causal vocabulary, I have no doubt.

11.6.4 Why Isn't Confounding a Statistical Concept?

In June 2001, I received two anonymous reviews of my paper "Causal Inference in the Health Sciences" (Pearl 2001c). The questions raised by the reviewers astounded me, for they reminded me of the archaic way some statisticians still think about causality and of the immense educational effort that still lies ahead. In the interest of contributing to this effort, I am including my reply in this chapter. Related discussion on the causal–statistical distinction is presented in Section 11.1.

Excerpts from Reviewers' Comments:

Reviewer 1.

"The contrast between statistical and causal concepts is overdrawn. Randomization, instrumental variables, and so forth have clear statistical definitions. … [the paper urges] 'that any systematic approach to causal analysis must require new mathematical notation.' This is false: there is a long tradition of informal – but systematic and successful – causal inference in the medical sciences."

Reviewer 2.

"The paper makes many sweeping comments which rely on distinguishing 'statistical' and 'causal' concepts … Also, included in the list of causal (and therefore, according to the paper, non-statistical) concepts is, for example, confounding, which is solidly founded in standard, frequentist statistics. Statisticians are inclined to say things like 'U is a potential confounder for examining the effect of treatment X on outcome Y when both U and X and U and Y are not independent. So why isn't confounding a statistical concept?' … If the author wants me to believe this, he's going to have to show at least one example of how the usual analyses fail."

---

[17] This inequality follows from (3.52) or (9.33). A complete characterization of coherence constraints is given in Tian, Kang, and Pearl (2006). As an example, for any three variables X, Y, Z, coherence dictates: $P(y \mid do(x, z)) - P(y, x \mid do(z)) - P(y, z \mid do(x)) + P(x, y, z) \geq 0$. If the structure of a causal graph is known, the conditions of Definition 1.3.1 constitute a complete characterization of all coherence requirements.

Tian and Pearl (2002c) and

Author's Second Reply:

The independence $Y_{xz} \perp\!\!\!\perp Z_{x*} \mid W$ actually holds in the graph shown in Figure 11.19.

This is because Y_{xz} is separated from Z_{x*} by the variable W_{x*}, in the "triple network" that you mentioned. The license to replace W with W_{x*} is obtained from rule 3 of *do*-calculus, which implies $W_{x*} = W$, since X is a nondescendant of X. This points to an important refinement needed in the twin network generalization: causal axioms may entail certain equality constraints among seemingly distinct counterfactual variables, and these hidden equalities need to be considered when we apply d-separation to counterfactual networks. A systematic way of encoding and managing these equalities is presented in Shpitser and Pearl (2007).

11.8 INSTRUMENTAL VARIABLES AND NONCOMPLIANCE

11.8.1 Tight Bounds under Noncompliance (Question to Author)

I am referring to the way you improved Manski's bounds on treatment effects when we have imperfect compliance. Which information does your approach exploit that the one by Manski does not? What is the intuition behind it?

Author's Reply:

We used the same information and same assumptions as Manski, and we derived the tight bounds using linear-programming analysis in the 16-dimensional space defined by the canonical partition of U (Balke and Pearl 1994a, 1995a). Manski, evidently, either did not aim at getting tight bounds, or was not aware of the power of partitioning U into its equivalence classes. Recall, this partition was unknown to economists before Frangakis and Rubin (2002) popularized it, under the rubric "principal stratification."

Manski's bounds, as I state on page 269, are tight under certain conditions, e.g., no contrarians. This means that one can get narrower bounds *only* when there are contrarians in the population, as in the examples discussed in Pearl (1995b). It is shown there how data representing the presence of contrarians can provide enough information to make the bounds collapse to a point. That article also gives an intuitive explanation of how this can happen.

It is important to mention at this point that the canonical partition conception, coupled with the linear programming method developed in Balke and Pearl (1994a, 1995a,b), has turned into a powerful analytical tool in a variety of applications. Tian and Pearl (2000) applied it to bound probabilities of causation; Kaufman et al. (2005) and Cai et al. (2008) used it to bound direct effects in the presence of confounded mediation, and, similarly, Imai et al. (2008) used it to bound natural direct and indirect effects. The closed-form expressions derived by this method enable researchers to assess what features of the distribution are critical for narrowing the widths of the bounds.

Rubin (2004), in an independent exploration, attempted to apply canonical partitions to the analysis of direct and indirect effects within the traditional potential-outcome framework but, lacking the graphical and structural perspectives, was led to conclude that such effects are "ill-defined" and "more deceptive than helpful." I believe readers of this book, guided by the structural roots of potential-outcome analysis, will reach more positive conclusions (see Sections 4.5 and 11.4.2).

and Sjölander (2009b)

and derive

$$P(y_{x'} \mid x) = \left[P(y_{x'}) - P(y, x')\right]/P(x). \quad = P(y \mid x') + [P(y_{x'}) - P(y \mid x)]/P(x).$$

In other words, $P(y_{x'} \mid x)$ is reducible to empirically estimable quantities; $P(y_{x'}) = P(y \mid do(x'))$ is estimable in experimental studies and the other quantities in observational studies. Moreover, if data support the equality $P(y_{x'}) = P(y, x')$, we can safely conclude that a treated patient would have zero chance of survival had the treatment not been taken. Those who mistrust counterfactual analysis a priori, as a calculus dealing with undefined quantities, would never enjoy the discovery that some of those quantities are empirically definable. Logic, when gracious, can rerun history for us.

The second puzzle was given intuitive explanation in the paragraph following equation (9.54).

The third puzzle is the one that gives most people a shock of disbelief. For a statistician, in particular, it is a rare case to be able to say anything certain about a specific individual who was not tested directly. This emanates from two factors. First, statisticians normally deal with finite samples, the variability of which rules out certainty in any claim, not merely about an individual but also about any property of the underlying distribution. This factor, however, should not enter into our discussion, for we have been assuming infinite samples throughout. (Readers should imagine that the numbers in Table 9.2 stand for millions.)

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

This argument, advanced by Freedman and Stark (1999), is incompatible with the way probability statements are used in ordinary discourse, for it implies that every probability statement about an individual must be a statement about a restricted subpopulation that shares *all* the individual's characteristics. Taken to the extreme, such a restrictive interpretation would insist on characterizing the plaintiff in minute detail, and would reduce PN to zero or one when all relevant details were accounted for. It is inconceivable that this interpretation underlies the intent of judicial standards. By using the wording "more probable than not," lawmakers have instructed us to ignore specific features for which data is not available, and to base our determination on the most specific features for which reliable data is available. In our example, two properties of Mr. A were noted: (1) that he died and (2) that he chose to use the drug; these were properly taken into account in bounding PN. If additional properties of Mr. A become known, and deemed relevant (e.g., that he had red hair, or was left-handed), these too could, in principle, be accounted for by restricting the analysis to data representing the appropriate subpopulations. However, in the absence of such data, and knowing in advance that we will never be able to match *all* the idiosyncratic properties of Mr. A, the lawmakers' specification must be interpreted relative to the properties at hand.

Lehmann, Dennis Lindley, Jacques A. Hagenaars, Jonathan Wilson, Stan Mulaik, Bill Shipley, Nozer D. Singpurwalla, Les Hayduk, Erich Battistin, Sampsa Hautaniemi, Melanie Wall, Susan Scott, Patrik Hoyer, Joseph Halpern, Phil Dawid, Sander Greenland, Arvid Sjölander, Eliezer S. Yudkowsky, UCLA students in CS262Z (Seminar in Causality, Spring 2006), and the UCLA Epidemiology Class – EPIDEM 200C.

I similarly thank all reviewers of the first edition of the book and the editors who helped bring these reviews to the attention of their readers. These include: *Choice* (Byerly 2000), *Structural Equation Modeling* (Shipley 2000a), *Chance* (McDonald 2001), *Technometrics* (Zelterman 2001), *Mathematical Reviews* (Lawry 2001), *Politische Vierteljahrsschrlft* (Didelez and Pigeot 2001), *Technological Forecasting & Social Change* (Payson 2001), *British Journal for the Philosophy of Science* (Gillies 2001), *Human Biology* (Chakraborty 2001), *The Philosophical Review* (Hitchcock 2001), *Intelligence* (O'Rourke 2001), *Journal of Marketing Research* (Rigdon 2002), *Tijdschrlft Voor* (Decock 2002), *Psychometrika* (McDonald 2002b), *International Statistical Review* (Lindley 2002), *Journal of Economic Methodology* (Leroy 2002), *Statistics in Medicine* (Didelez 2002), *Journal of Economic Literature* (Swanson 2002), *Journal of Mathematical Psychology* (Butler 2002), *IIE Transactions* (Gursoy 2002), *Royal Economic Society* (Hoover 2003), *Econometric Theory* (Neuberg 2003), *Economica* (Abbring 2003), *Economics and Philosophy* (Woodward 2003), *Sociological Methods and Research* (Morgan 2004), *Review of Social Economy* (Boumans 2004), *Journal of the American Statistical Association* (Hadlock 2005), and *Artificial Intelligence* (Kyburg 2005).

Thanks also go to the contributors to UCLA's *Causality* blog (http://www.mii.ucla.edu/causality/) and to William Hsu, the blog's curator.

A special word of appreciation and admiration goes to Dennis Lindley, who went to the trouble of studying my ideas from first principles, and allowed me to conclude that readers from the statistics-based sciences would benefit from this book. I am fortunate that our paths have crossed and that I was able to witness the intellect, curiosity, and integrity of a true gentleman.

This chapter could not have been written without the support and insight of Jin Tian, Avin Chen, Carlo Brito, Blai Bonet, Mark Hopkins, Ilya Shpitser, Azaria Paz, Manabu Kuroki, Zhihong Cai, Kaoru Mulvihill, and all members of the Cognitive System Laboratory at UCLA who, in the past six years, continued to explore the green pastures of causation while I was summoned to mend a world that took the life of my son Daniel (murdered by extremists in Karachi, Pakistan, 2002). These years have convinced me, beyond a shadow of doubt, that the forces of reason and enlightenment will win over fanaticism and inhumanity.

Finally, I dedicate this edition to my wife, Ruth, for her strength, love, and guidance throughout our ordeal, to my daughters, Tamara and Michelle, and my grandchildren, Leora, Torri, Adam, Ari, and Evan for being with me through this journey and making it meaningful and purposeful.

Hitchcock, 1995 C. Hitchcock. The mishap of Reichenbach's fall: Singular vs. general causation. *Philosophical Studies,* 78:257–291, 1995.

Hitchcock, 1996 C.R. Hitchcock. Causal decision theory and decision theoretic causation. *Nous,* 30(4):508–526, 1996.

Hitchcock, 1997 C. Hitchcock. Causation, probabilistic, 1997. In *Stanford Encyclopedia of Philosophy,* online at: http://plato.stanford.edu/entries/causation-probabilistic.

Hitchcock, 2001 C. Hitchcock. Book reviews: Causality: Models, Reasoning, and Inference. *The Philosophical Review,* 110(4):639–641, 2001.

Hitchcock, 2007 C.R. Hitchcock. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review,* 116:495–532, 2007.

Hitchcock, 2008 C.R. Hitchcock. Structural equations and causation: Six counterexamples. *Philosophical Studies,* page DOI 10.1007/s 11098–008–9216–2, 2008.

Hoel et al., 1971 P.G. Hoel, S.C. Port, and C.J. Stone. *Introduction to Probability Theory*. Houghton Mifflin Company, Boston, 1971.

Holland and Rubin, 1983 P.W. Holland and D.B. Rubin. On Lord's paradox. In H. Wainer and S. Messick, editors, *Principals of Modern Psychological Measurement,* pages 3–25. Lawrence Earlbaum, Hillsdale, NJ, 1983.

Holland, 1986 P.W. Holland. Statistics and causal inference. *Journal of the American Statistical Association,* 81(396):945–960, December 1986.

Holland, 1988 P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology,* pages 449–484. American Sociological Association, Washington, D.C., 1988.

Holland, 1995 P.W. Holland. Some reflections on Freedman's critiques. *Foundations of Science,* 1:50–57, 1995.

Holland, 2001 P.W. Holland. The false linking of race and causality: Lessons from standardized testing. *Race and Society,* 4(2): 219–233, 2001.

Hoover, 1990 K.D. Hoover. The logic of causal inference. *Economics and Philosophy,* 6:207–234, 1990.

Hoover, 2001 K. Hoover. *Causality in Macroeconomics*. Cambridge University Press, New York, 2001.

Hoover, 2003 K.D. Hoover. Book reviews: Causality: Models, Reasoning, and Inference. *Economic Journal,* 113:F411–F413, 2003.

Hoover, 2004 K.D. Hoover. Lost causes. *Journal of the History of Economic Thought,* 26(2):149–164, June 2004.

Hoover, 2008 K.D. Hoover. Causality in economics and econometrics. In S.N. Durlauf and L.E. Blume, editors, *From The New Palgrave Dictionary of Economics*. Palgrave Macmillan, New York, NY, 2nd edition, 2008.

Hopkins and Pearl, 2002 M. Hopkins and J Pearl. Strategies for determining causes of events. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence,* pages 546–552. AAAI Press/ The MIT Press, Menlo Park, CA, 2002.

Howard and Matheson, 1981 R.A. Howard and J.E. Matheson. Influence diagrams. *Principles and Applications of Decision Analysis,* 1981. Strategic Decisions Group, Menlo Park, CA. Reprinted in *Decision Analysis* 2(3): 129–143, 2005.

Howard, 1960 R.A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

Howard, 1990 R.A. Howard. From influence to relevance to knowledge. In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets, and Decision Analysis,* pages 3–23. Wiley and Sons, Ltd., New York, NY, 1990.

Hoyer et al., 2006 P. Hoyer, S. Shimizu, and A.J. Kerminen. Estimation of linear, non-Gaussian causal models in presence of confounding latent variables. In *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM'06)*, pages 155–162. Institute of Information Theory and Automation, Prague, Czech Republic, 2006.

Huang and Valtorta, 2006a Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence,* pages 217–224. AUAI Press, Corvallis, OR, 2006.

Huang and Valtorta, 2006b Y. Huang and M. Valtorta. Identifiability in causal Bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence,* pages 1149–1154, AAAI Press, Menlo Park, CA, July 2006.

Suppes, 1970 P. Suppes. *A Probabilistic Theory of Causality.* North-Holland Publishing Co., Amsterdam, 1970.

Suppes, 1988 P. Suppes. Probabilistic causality in space and time. In B. Skyrms and W.L. Harper, editors, *Causation, Chance, and Credence.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

Swanson and Granger, 1997 N.R. Swanson and C.W.J. Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92:357–367, 1997.

Swanson, 2002 N.R. Swanson. Book reviews: Causality: Models, Reasoning, and Inference. *Journal of Economic Literature,* XL:925–926, 2002.

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

Tian and Pearl, 2001a J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* pages 512–521. Morgan Kaufmann, San Francisco, CA, 2001.

Tian and Pearl, 2001b J. Tian and J. Pearl. Causal discovery from changes: A Bayesian approach. Technical Report R-285, Computer Science Department, UCLA, February 2001.

Tian and Pearl, 2002a J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence,* pages 567–573. AAAI Press/ The MIT Press, Menlo Park, CA, 2002.

Tian and Pearl, 2002b J. Tian and J Pearl. On the testable implications of causal models with hidden variables. In A. Darwiche and N. Friedman, editors, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 519–527. Morgan Kaufmann, San Francisco, CA, 2002.

Tian et al., 1998 J. Tian, A. Paz, and J. Pearl. Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA, 1998.

Tian et al., 2006 J. Tian, C. Kang, and J. Pearl. A characterization of interventional distributions in semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence,* pages 1239–1244. AAAI Press, Menlo Park, CA, 2006.

Tversky and Kahneman, 1980 A. Tversky and D. Kahneman. Causal schemas in judgments under uncertainty. In M. Fishbein, editor, *Progress in Social Psychology,* pages 49–72. Lawrence Erlbaum, Hillsdale, NJ, 1980.

VanderWeele and Robins, 2007 T.J. VanderWeele and J.M. Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology,* 18(5):561–568, 2007.

Verma and Pearl, 1988 T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence,* pages 352–359, Mountain View, CA, 1988. Also in R. Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4,* Elsevier Science Publishers, 69–76, 1990.

Verma and Pearl, 1990 T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence,* pages 220–227, Cambridge, MA, July 1990. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6,* Elsevier Science Publishers, B.V, 255–268, 1991.

Verma and Pearl, 1992 T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In D. Dubois, M.P. Wellman, B. D'Ambrosio, and P. Smets, editors, *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence,* pages 323–330. Morgan Kaufmann, Stanford, CA, 1992.

Verma, 1993 T.S. Verma. Graphical aspects of causal models. Technical Report R-191, UCLA, Computer Science Department, 1993.

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

Wang et al., 2009 X. Wang, Z. Geng, H. Chen, and X. Xie. Detecting multiple confounders. *Journal of Statistical Planning and Inference,* 139: 1073–1081, 2009.

Wasserman, 2004 L. Wasserman. *All of Statistics*: *A Concise Course in Statistical Inference*. Springer Science+Business Media, Inc., New York, NY, 2004.

Tian and Pearl, 2002c J. Tian and J. Pearl. A New Characterization of the Experimental Implications of Causal Bayesian Networks. *Proceedings of the Eighteenth National Conference on Artificial Intelligence,* pages 574-579. AAAI Press/TheMIT Press: Menlo Park, CA 2002.

Tian and Pearl, 2003 J. Tian and J. Pearl. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, CA, 2003.