To appear in P. McKay Illari, F. Russo, and J. Williamson (Eds.),
Causality in the Sciences, Clarendon Press, Oxford, 2009.
(Chapter and page numbering still to be determined.)

TECHNICAL REPORT
R-351
July 2009

1

THE STRUCTURAL THEORY OF CAUSATION

JUDEA PEARL

Abstract

This paper presents a general theory of causation based on the Structural Causal Model (SCM) described in (Pearl, 2000*a*). The theory subsumes and unifies current approaches to causation, including graphical, potential outcome, probabilistic, decision analytical, and structural equation models, and provides both a mathematical foundation and a friendly calculus for the analysis of causes and counterfactuals. In particular, the paper demonstrates how the theory engenders a coherent methodology for inferring (from a combination of data and assumptions) answers to three types of causal queries: (1) queries about the effects of potential interventions, (2) queries about probabilities of counterfactuals, and (3) queries about direct and indirect effects.

Keywords: Structural equation models, confounding, graphical methods, counterfactuals, causal effects, potential-outcome, probabilistic causation.

1.1 Introduction

The research questions that motivate most studies in the health, social and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from distributions alone.

Any conception of causation worthy of the title "theory" must be able to (1) represent these questions in a formal language, (2) provide a precise language for communicating assumptions under which the questions need to be answered, (3) provide a systematic way of answering at least some of these questions and labeling others "unanswerable," and (4) provide a method of determining what assumptions or new measurements would be needed to answer the "unanswerable" questions.[1]

[1] For example, a theory may conclude that the information at hand is not sufficient for determining the efficacy of a drug unless certain assumptions are deemed plausible, or unless data from a specific experimental study were made available. Such conclusion constitutes a valid "solution," provided no better solution exists.

A "general theory" of causation should do more. In addition to embracing *all* questions judged to have causal character, a general theory must also *subsume* any other theory or method that scientists have found useful in exploring the various aspects of causation, be they epistemic, methodological or practical. In other words, any alternative theory need to evolve as a special case of the "general theory" when restrictions are imposed on either the model, the type of assumptions admitted, or the language in which those assumptions are cast.

This paper presents a theory of causation that satisfies the criteria above. It is based on the Structural Causal Model (SCM) developed in (Pearl, 1995, 2000*a*) which combines features of the structural equation models (SEM) used in economics (Haavelmo, 1943) and social science (Duncan, 1975), the potential-outcome notation of Neyman (1923) and Rubin (1974), and the graphical models developed for probabilistic reasoning (Pearl, 1988; Lauritzen, 1996) and causal analysis (Spirtes *et al.*, 2000; Pearl, 2000*a*). The theory presented forms a coherent whole that supercedes the sum of its parts.

Although the basic elements of SCM were introduced in the mid 1990's (Pearl, 1995), and have been adapted warmly by epidemiologists (Greenland *et al.*, 1999; Glymour and Greenland, 2008), statisticians (Cox and Wermuth, 2004; Lauritzen, 2001), and social scientists (Morgan and Winship, 2007), its potentials as a comprehensive theory of causation are yet to be fully utilized. Some have congratulated the SCM for generalizing econometric models from linear to non-parametric analysis (Heckman, 2008), some have marveled at the clarity and transparency of the graphical representation (Greenland and Brumback, 2002), others praised the flexibility of the $do(x)$ operator (Hitchcock, 2001; Lindley, 2002; Woodward, 2003) and, naturally, many have used the SCM to weed out myths and misconceptions from outdated traditions (Meek and Glymour, 1994; Greenland *et al.*, 1999; Cole and Hernán, 2002; Arah, 2008; Shrier, 2009; Pearl, 2009*b*) Still, the more profound contributions of SCM, those stemming from its role as a comprehensive theory of causation, have not been fully explicated. These include:

1. The unification of the graphical, potential outcome, structural equations, decision analytical (Dawid, 2002), interventional (Woodward, 2003), sufficient component (Rothman, 1976) and probabilistic approaches to causation; with each approach viewed as a restricted special aspect of the SCM.
2. The axiomatization and algorithmization of counterfactual expressions.
3. Defining and identifying joint probabilities of counterfactual statements.
4. Reducing the evaluation of actions and policies to algorithmic level of analysis.
5. Solidifying the mathematical foundations of the potential-outcome model, and formulating the counterfactual foundations of structural equation models.
6. Demystifying enigmatic notions such as "confounding," "ignorability," "exchangeability," "superexogeneity" and others, which have emerged from

"black-box" approaches to causation.

7. Providing a transparent language for communicating causal assumptions and defining causal problems.

This paper presents the main features of the structural theory by, first, contrasting causal analysis with standard statistical analysis (Section 1.2), second, presenting a friendly formalism for counterfactual analysis, within which most (if not all) causal questions can be formulated and resolved (Section 1.3 and 1.4) and, finally, contrasting the structural theory with two other frameworks: probabilistic causation (Section 1.5) and the Neyman-Rubin potential-outcome model (Section 1.6). The analysis will be demonstrated by attending to three types of queries: (1) queries about the effect of potential interventions, (Section 1.3.1 and 1.3.2), (2) queries about counterfactuals (Section 1.3.3) and (3) queries about direct and indirect effects (Section 1.4).

1.2 From Associational to Causal Analysis: Distinctions and Barriers

1.2.1 *The Basic Distinction: Coping With Change*

The aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or external interventions.

This distinction implies that causal and associational concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions which identify relationships that remain invariant when external conditions change.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: one cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.[2]

[2]The methodology of "causal discovery" (Spirtes *et al.* 2000; Pearl 2000*a*, Chapter 2) is likewise based on the causal assumption of "faithfulness" or "stability."

1.2.2 *Formulating the Basic Distinction*

A useful demarcation line that makes the distinction between associational and causal concepts crisp and easy to apply, can be formulated as follows. An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odd ratio, marginalization, conditionalization, "controlling for," and so on. Examples of causal concepts are: randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution, and so on. The former can, while the latter cannot be defined in term of distribution functions.

This demarcation line is extremely useful in causal analysis for it helps investigators to trace the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must rely on some premises that invoke such concepts; it cannot be inferred from, or even defined in terms statistical associations alone.

1.2.3 *Ramifications of the Basic Distinction*

This principle has far reaching consequences that are not generally recognized in the standard statistical literature. Many researchers, for example, are still convinced that confounding is solidly founded in standard, frequentist statistics, and that it can be given an associational definition saying (roughly): "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." That this definition and all its many variants must fail (Pearl, 2000*a*, Section 6.2)[3] is obvious from the demarcation line above; if confounding were definable in terms of statistical associations, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders and obtain unbiased estimates of causal effects. This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies. Hence the definition must be false. Therefore, to the bitter disappointment of generations of epidemiologist and social science researchers, confounding bias cannot be detected or corrected by statistical methods alone; one must make some judgmental assumptions regarding causal relationships in the problem before an adjustment (e.g., by stratification) can safely correct for confounding bias.

Another ramification of the sharp distinction between associational and causal concepts is that any mathematical approach to causal analysis must acquire new notation for expressing causal relations – probability calculus is insufficient. To illustrate, the syntax of probability calculus does not permit us to express the

[3]For example, any intermediate variable U on a causal path from X to Y satisfies this definition, without confounding the effect of X on Y.

simple fact that "symptoms do not cause diseases", let alone draw mathematical conclusions from such facts. All we can say is that two events are dependent—meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease|symptom)$ from causal dependence, for which we have no expression in standard probability calculus. Scientists seeking to express causal relationships must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease."

1.2.4 *Two Mental Barriers: Untested Assumptions and New Notation*

The preceding two requirements: (1) to commence causal analysis with untested,[4] theoretically or judgmentally based assumptions, and (2) to extend the syntax of probability calculus, constitute the two main obstacles to the acceptance of causal analysis among statisticians and among professionals with traditional training in statistics.

Associational assumptions, even untested, are testable in principle, given sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to prior causal assumptions, say that treatment does not change gender, remains substantial regardless of sample size.

This makes it doubly important that the notation we use for expressing causal assumptions be meaningful and unambiguous so that one can clearly judge the plausibility or inevitability of the assumptions articulated. Statisticians can no longer ignore the mental representation in which scientists store experiential knowledge, since it is this representation, and the language used to access it that determine the reliability of the judgments upon which the analysis so crucially depends.

How does one recognize causal expressions in the statistical literature? Those versed in the potential-outcome notation (Neyman, 1923; Rubin, 1974; Holland, 1988), can recognize such expressions through the subscripts that are attached to counterfactual events and variables, e.g. $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions, e.g. $Y(0)$, $Y(1)$, $Y(x,u)$ or $Z(x,y)$.) The expression $Y_x(u)$, for example, stands for the value that outcome Y would take in individual u, had treatment X been at level x. If u is chosen at random, Y_x is a random variable, and one can talk about the probability that Y_x would attain a value y in the population, written $P(Y_x = y)$ (see Section 1.6 for semantics). Alternatively, (Pearl, 1995) used expressions of the form $P(Y = y|set(X = x))$ or $P(Y = y|do(X = x))$ to denote the probability (or frequency) that event

[4] By "untested" I mean untested using frequency data in nonexperimental studies.

$(Y = y)$ would occur if treatment condition $X = x$ were enforced uniformly over the population.[5] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality.[6]

However, few have taken seriously the textbook requirement that any introduction of new notation must entail a systematic definition of the syntax and semantics that governs the notation. Moreover, in the bulk of the statistical literature before 2000, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations, and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate not be affected by a treatment, a necessary assumption for the control of confounding (Cox, 1958, p. 48), is expressed in plain English, not in a mathematical expression.

Remarkably, though the necessity of explicit causal notation is now recognized by many academic scholars, the use of such notation has remained enigmatic to most rank and file researchers, and its potentials still lay grossly underutilized in the statistics based sciences. The reason for this, can be traced to the unfriendly semi-formal way in which causal analysis has been presented to the research community, resting primarily on the restricted paradigm of controlled randomized trials advanced by (Rubin, 1974).

The next section provides a conceptualization that overcomes these mental barriers; it offers both a friendly mathematical machinery for cause-effect analysis and a formal foundation for counterfactual analysis.

1.3 Structural Causal Models (SCM) and The Language of Diagrams

1.3.1 *Semantics: Causal Effects and Counterfactuals*

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920's by the geneticist Sewall Wright (1921), who used a combination of equations and graphs. For example, if X stands for a disease variable and Y stands for a certain symptom of the disease, Wright would write a linear equation:

$$y = \beta x + u_Y \tag{1.1}$$

where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u_Y stands for all factors, other than the disease in question, that could possibly affect Y. In interpreting this equation one should

[5] Clearly, $P(Y = y|do(X = x))$ is equivalent to $P(Y_x = y)$, This is what we normally assess in a controlled experiment, with X randomized, in which the distribution of Y is estimated for each level x of X.

[6] These notational clues should be useful for detecting inadequate definitions of causal concepts; any definition of confounding, randomization or instrumental variables that is cast in standard probability expressions, void of graphs, counterfactual subscripts or $do(*)$ operators, can safely be discarded as inadequate.

think of a physical process whereby Nature *examines* the values of x and u_Y and, accordingly, *assigns* variable Y the value $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease X, one could write $x = u_X$, where U_X stands for all factors affecting X.

To express the directionality inherent in this process, Wright augmented the equation with a diagram, later called "path diagram," in which arrows are drawn from (perceived) causes to their (perceived) effects and, more importantly, the absence of an arrow makes the empirical claim that Nature assigns values to one variable while ignoring the other. In Figure 1.1, for example, the absence of arrow from Y to X represent the claim that symptom Y is not among the factors U_X which affect disease X.

The variables U_X and U_Y are called "exogenous"; they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model.

If correlation is judged possible between two exogenous variables, U_Y and U_X, it is customary to connect them by a dashed double arrow, as shown in Figure 1.1(b).

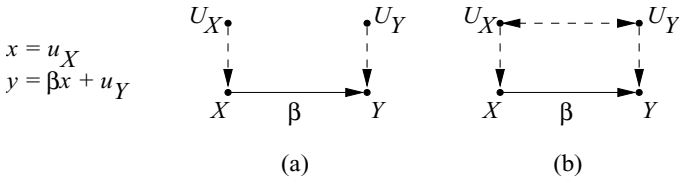

FIG. 1.1. A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

To summarize, path diagrams encode causal assumptions via missing arrows, representing claims of zero influence, and missing double arrows (e.g., between U_X and U_Y), representing the (causal) assumption $Cov(U_Y, U_X)$=0.

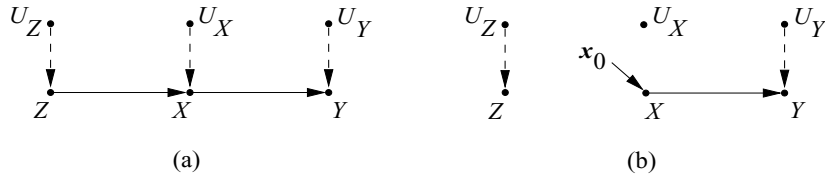

FIG. 1.2. (a) The diagram associated with the structural model of equation (1.2). (b) The diagram associated with the modified model, M_{x_0}, of equation (1.3), representing the intervention $do(X = x_0)$.

The generalization to nonlinear systems of equations is straightforward. For

example, the non-parametric interpretation of the diagram of Figure 1.2(a) corresponds to a set of three functions, each corresponding to one of the observed variables:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y) \end{aligned} \tag{1.2}$$

where U_Z, U_X and U_Y are assumed to be jointly independent but, otherwise, arbitrarily distributed.

Remarkably, unknown to most economists and pre-2000 philosophers,[7] structural equation models provide a formal interpretation and symbolic machinery for analyzing counterfactual relationships of the type: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. Here U represents the vector of all exogenous variables.[8]

The key idea is to interpret the phrase "had X been x_0" as an instruction to modify the original model and replace the equation for X by a constant x_0, yielding the sub-model.

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_Y) \end{aligned} \tag{1.3}$$

the graphical description of which is shown in Figure 1.2(b).

This replacement permits the constant x_0 to differ from the actual value of X (namely $f_X(z, u_X)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multi-stage models, where the dependent variable in one equation may be an independent variable in another (Balke and Pearl, 1994*ab*; Pearl 2000*b*). For example, to compute $E(Y_{x_0})$, the expected effect of *setting* X to x_0, (also called the *average causal effect* of X on Y, denoted $E(Y|do(x_0))$ or, generically, $E(Y|do(x))$), we solve equation (1.3) for Y in terms of the exogenous variables, yielding $Y_{x_0} = f_Y(x_0, u_Y)$, and average over U_Y. It is easy to show that in this simple system, the answer can be obtained without knowing the form of the function $f_Y(x, u_Y)$ or the distribution $P(u_Y)$. The answer is given by:

$$E(Y_{x_0}) = E(Y|do(X = x_0) = E(Y|x_0)$$

[7]Connections between structural equations and a restricted class of counterfactuals were recognized by (Simon and Rescher, 1966). These were generalized by (Balke and Pearl, 1995) who used modified models to permit counterfactual conditioning on dependent variables. This development seems to have escaped Collins *et al.* (2004).

[8]Because $U = u$ may contain detailed information about a situation or an individual, $Y_x(u)$ is related to what philosophers called "token causation," while $P(Y_x = y|Z = z)$ characterizes "Type causation," that is, the tendency of X to influence Y in a sub-population characterized by $Z = z$.

which is computable from the distribution $P(x, y, z)$, hence estimable from observed samples of $P(x, y, z)$. This result hinges on the assumption that U_Z, U_X, and U_Y are mutually independent and on the topology of the graph (e.g., that there is no direct arrow from Z to Y.)

In general, it can be shown (Pearl, 2000*a*, Chapter 3) that, whenever the graph is Markovian (i.e., acyclic with independent exogenous variables) the post-interventional distribution $P(Y = y|do(X = x))$ is given by the following expression:

$$P(Y = y|do(X = x)) = \sum_t P(y|t, x)P(t) \tag{1.4}$$

where T is the set of direct causes of X (also called "parents") in the graph. Again, we see that all factors on the right hand side are estimable from the distribution P of observed variables and, hence, the counterfactual probability $P(Y_x = y)$ is estimable with mere partial knowledge of the generating process – the topology of the graph and independence of the exogenous variables is all that is needed.

When some variables in the graph (e.g., the parents of X) are unobserved, we may not be able to learn (or "identify" as it is called) the post-intervention distribution $P(y|do(x))$ by simple conditioning, and more sophisticated methods would be required. Likewise, when the query of interest involves several hypothetical worlds simultaneously, e.g., $P(Y_x = y, Y_{x'} = y')$[9], the Markovian assumption may not suffice for identification and additional assumptions, touching on the form of the data-generating functions (e.g., monotonicity) may need to be invoked. These issues will be discussed in Sections 1.3.3 and 1.6.

This interpretation of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation models, used in economics and social science and the Neyman-Rubin potential-outcome framework to be discussed in Section 1.6. But first we discuss two long-standing problems that have been completely resolved in purely graphical terms, without delving into algebraic techniques.

1.3.2 *Confounding and Causal Effect Estimation*

The central target of most studies in the social and health sciences is the elucidation of cause-effect relationships among variables of interests, for example, treatments, policies, preconditions and outcomes. While good statisticians have always known that the elucidation of causal relationships from observational studies must be shaped by assumptions about how the data were generated, the relative roles of assumptions and data, and ways of using those assumptions to eliminate confounding bias have been a subject of much controversy. The structural framework of Section 1.3.1 puts these controversies to rest.

Covariate Selection: The back-door criterion Consider an observational study where we wish to find the effect of X on Y, for example, treatment on response,

[9]Read: The probability that Y would be y if X were x and y' if X were x'.

and assume that the factors deemed relevant to the problem are structured as in Figure 1.3; some are affecting the response, some are affecting the treatment

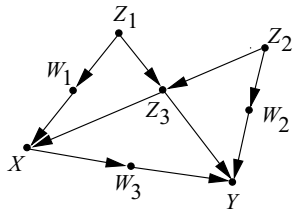

FIG. 1.3. Graphical model illustrating the back-door criterion for identifying the causal effect of X on Y. Error terms are not shown explicitly.

and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or life style, others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set" or a set "appropriate for adjustment". The problem of defining a sufficient set, let alone finding one, has baffled epidemiologists and social science for decades (see (Greenland *et al.*, 1999; Pearl, 1998; Pearl, 2003) for review).

The following criterion, named "back-door" in (Pearl, 1993*a*), settles this problem by providing a graphical method of selecting a sufficient set of factors for adjustment. It states that a set S is appropriate for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.[10]

Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained above, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The implication of finding a sufficient set S is that, stratifying on S is guaranteed to remove all confounding bias relative the causal effect of X on Y. In other words, it renders the causal effect of X on Y estimable, via

[10] A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. See (Pearl, 2000*a*, pp. 16-7). If S blocks *all* paths from X to Y it is said to "d-separate X and Y."

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \tag{1.5}$$

Since all factors on the right hand side of the equation are estimable (e.g., by regression) from the pre-interventional data, the causal effect can likewise be estimated from such data without bias.

The back-door criterion allows us to write equation (1.5) directly, after selecting a sufficient set S from the diagram, without resorting to any algebraic manipulation. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether "X is conditionally ignorable given S," a formidable mental task required in the potential-outcome framework (Rosenbaum and Rubin, 1983). The criterion also enables the analyst to search for an optimal set of covariate—namely, a set S that minimizes measurement cost or sampling variability (Tian *et al.*, 1998).

General control of confounding Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. A much more general identification criterion is provided by the following theorem:

Theorem 1.1 (*Tian and Pearl 2002*)
A sufficient condition for identifying the causal effect $P(y|do(x))$ is that every path between X and any of its children traces at least one arrow emanating from a measured variable.[11]

For example, if W_3 is the only observed covariate in the model of Figure 1.3, then there exists no sufficient set for adjustment (because no set of observed covariates can block the paths from X to Y through Z_3), yet $P(y|do(x))$ can nevertheless be estimated since every path from X to W_3 (the only child of X) traces either the arrow $X \to W_3$, or the arrow $W_3 \to Y$, both emanating from a measured variable (W_3). In this example, the variable W_3 acts as a "mediating instrumental variable" (Pearl, 1993*b*; Chalak and White, 2006) and yields the estimand:

$$\begin{aligned} P(Y = y|do(X = x)) &= \sum_{w_3} P(W_3 = w_3|do(X = x))P(Y = y|do(W_3 = w_3)) \\ &= \sum_{w_3} P(w_3|x) \sum_{x'} P(y|w_3, x')P(x') \end{aligned} \tag{1.6}$$

More recent results extend this theorem by (1) presenting a necessary and sufficient condition for identification (Shpitser and Pearl, 2006*a*), and (2) extending the condition from causal effects to any counterfactual expression (Shpitser and

[11] Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.

Pearl, 2007). The corresponding unbiased estimands for these causal quantities are readable directly from the diagram.

The mathematical derivation of causal effect estimands, like equations (1.5) and (1.6) is merely a first step toward computing quantitative estimates of those effects from finite samples, using the rich traditions of statistical estimation and machine learning. Although the estimands derived in (1.5) and (1.6) are non-parametric, this does not mean 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 (1.6) can be converted to the product $E(Y|do(x)) = r_{W_3X} r_{YW_3 \cdot X} x$, where $r_{YZ \cdot X}$ is the (standardized) coefficient of Z in the regression of Y on Z and X. More sophisticated estimation techniques can be found in (Rosenbaum and Rubin, 1983), and (Robins, 1999). For example, the "propensity score" method of (Rosenbaum and Rubin, 1983) was found to be quite useful when the dimensionality of the adjusted covariates is high and the data is sparse (see Pearl (2009*a*, pp. 348-42)).

It should be emphasized, however, that contrary to conventional wisdom (e.g., (Rubin, 2009)), propensity score methods are merely efficient estimators of the right hand side of (1.5); they cannot be expected to reduce bias in case the set S does not satisfy the back-door criterion (Pearl, 2009*abc*).

1.3.3 *Counterfactual Analysis in Structural Models*

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, in much the same way that not all causal questions can be answered from experimental studies. For example, questions of *attribution* (also called "causes of effects" (Dawid, 2000), e.g., I took an aspirin and my headache is gone, was it *due* to the aspirin?) or of *susceptibility* (e.g., I am a healthy non-smoker, would I be as healthy had I been a smoker?) cannot be answered from experimental studies, and naturally, this kind of questions cannot be expressed in $P(y|do(x))$ notation.[12] To answer such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$.

As noted in Section 1.3.1, the structural definition of counterfactuals involves modified models, like M_{x_0} of equation (1.3), formed by the intervention $do(X = x_0)$ (Figure 1.2(b)). Call the solution of Y in model M_x the *potential response* of Y to x, and denote it by the symbol $Y_x(u)$. In general, then, the formal definition of the counterfactual $Y_x(u)$ in SCM is given by (Pearl, 2000*a*, p. 98):

$$Y_x(u) = Y_{M_x}(u). \tag{1.7}$$

[12] The reason for this fundamental limitation is that no death case can be tested twice, with and without treatment. For example, if we measure equal proportions of deaths in the treatment and control groups, we cannot tell how many death cases are actually attributable to the treatment itself; it is quite possible that many of those who died under treatment would be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

The quantity $Y_x(u)$ can be given experimental interpretation; it stands for the way an individual with characteristics (u) would respond, had the treatment been x, rather than the treatment $x = f_X(u)$ actually received by that individual. In our example, since Y does not depend on v and w, we can write:

$$Y_{x_0}(u_Y, u_X, u_Z) = Y_{x_0}(u_Y) = f_Y(x_0, u_Y).$$

Clearly, the distribution $P(u_Y, u_X, u_Z)$ induces a well defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as '$Y_{x_0} = y$ AND $Y_{x_1} = y'$,' which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming linear equations (as in Figure 1.1),

$$x = u_X, \qquad y = \beta x + u_Y,$$

the conditions $Y = y_0$ and $X = x_0$ yield $u_X = x_0$ and $u_Y = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{x_1} must take on the value: $Y_{x_1} = \beta x_1 + u_Y = \beta(x_1 - x_0) + y_0$. In other words, if X were x_1 instead of x_0, Y would increase by β times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of U and, for that reason, attributional queries are generally not identifiable in nonparametric models (Pearl, 2000*a*, Chapter 9).

In general, if x and x' are incompatible then Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$."[13] Such concerns have been a source of objections to treating counterfactuals as jointly distributed random variables (Dawid, 2000). The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels neutralizes these objections (Pearl, 2000*b*), since the contradictory joint statement is mapped into an ordinary event, one where the background variables satisfy both statements simultaneously, each in its own distinct submodel; such events have well defined probabilities.

The structural interpretation of counterfactuals also provides the conceptual and formal basis for the Neyman-Rubin potential-outcome framework, an approach to causation that takes a controlled randomized trial (CRT) as its starting paradigm, assuming that nothing is known to the experimenter about the science behind the data. This "black-box" approach, which has thus far been denied the benefits of graphical or structural analyses, was developed by statisticians who found it difficult to cross the two mental barriers discussed in Section 1.2.4. Section 1.6 establishes the precise relationship between the structural and potential-outcome paradigms, and outlines how the latter can benefit from the richer representational power of the former.

[13] For example, "The probability is 80% that Joe belongs to the class of patients who will be cured if they take the drug and die otherwise."

1.4 Mediation: Direct and Indirect Effects

1.4.1 *Direct versus Total Effects:*

The causal effect we have analyzed so far, $P(y|do(x))$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

Another example concerns the identification of neural pathways in the brain or the structural features of protein-signaling networks in molecular biology (Brent and Lok, 2005). Here, the decomposition of effects into their direct and indirect components carries theoretical scientific importance, for it predicts behavior under a rich variety of hypothetical interventions.

In all such examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between gender (X) and hiring (Y) for a given level of qualification Z, because, by conditioning on the mediator Z, we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \rightarrow Z \leftarrow U \rightarrow Y$, where X has no direct effect on Y. Physically holding Z constant should eliminate the association between X and Y, as can be seen by deleting all arrows entering Z. But if we were to condition on Z, a spurious association would be created through U (unobserved) that might be construed as a direct effect of X on Y.

Using the $do(x)$ notation, and focusing on differences of expectations, this leads to a simple definition of *controlled direct effect*:

$$CDE \triangleq E(Y|do(x'), do(z)) - E(Y|do(x), do(z))$$

or, equivalently, using counterfactual notation:

$$CDE \triangleq E(Y_{x'z}) - E(Y_{xz})$$

where Z is any set of mediating variables that intercept all indirect paths between X and Y. Graphical identification conditions for expressions of the type

$E(Y|do(x), do(z_1), do(z_2), \ldots, do(z_k))$ were derived by (Pearl and Robins, 1995) (see Pearl (2000*a*, Chapter 4)) and invoke sequential application of the back-door conditions discussed in Section 1.3.2.

1.4.2 *Natural Direct Effects*

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y; therefore, the direct effect is independent of the values at which we hold Z. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z).

When the direct effect is sensitive to the levels at which we hold Z, it is often meaningful to average the direct effect over those levels. 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, 1991) called "pure" and (Pearl, 2001) called "natural," mirrors 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)).

Extending the subscript notation to express nested counterfactuals (Pearl, 2001) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{1.8}$$

Here, 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 original 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 footnote 12 for intuitive explanation). (Pearl, 2001) has nevertheless shown that, if certain assumptions of "no confounding" are deemed valid,[14] the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y|do(x', z)) - E(Y|do(x, z))]P(z|do(x)). \tag{1.9}$$

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z|do(x))$ as a weighing function.

[14] One sufficient condition is that $Z_x \perp\!\!\!\perp Y_{x',z}|W$ holds for some set W of measured covariates. See details and graphical criteria in (Pearl, 2001; Pearl, 2005) and in (Petersen *et al.*, 2006).

In particular, expression (1.9) is both valid and identifiable in Markovian models, where each term on the right can be reduced to a "*do*-free" expression using equation (1.4).

1.4.3 *Natural Indirect Effects*

Remarkably, the definition of the natural direct effect (1.8) 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, 2001):

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{1.10}$$

which is almost identical to the direct effect (equation (1.8)) 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,

$$TE_{x,x'}(Y) \triangleq E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{1.11}$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{1.12}$$

Since each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.

Note that, although it cannot be expressed in *do*-notation, the indirect effect has clear policy-making implications. For example: in the hiring discrimination context, 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.

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, 2001) 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. The mantra "No causation without manipulation" must be rejected. (See Pearl (2000*a*, Section 11.4.5.),

It is remarkable that counterfactual quantities like DE and ID 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 characterization of those conditions is given in (Shpitser and Pearl, 2007).

Additional examples of this "marvel of formal analysis" are given in (Pearl, 2000*a*, Chapters 7, 9, 11). It constitutes an unassailable argument in defense of counterfactual analysis, as expressed in (Pearl, 2000*b*) against the stance of (Dawid, 2000).

1.5 Structural versus Probabilistic Causality

Probabilistic causality (PC) is a branch of philosophy that has attempted, for the past several decades, to characterize the relationship between cause and effect using the tools of probability theory (Hitchcock, 2003; Williamson, ming). Our discussion of Section 1.2 rules out any such characterization and, not surprisingly, the PC program is known mainly for the difficulties it has encountered, rather than its achievements. This section explains the main obstacle that has kept PC at bay for over half a century, and demonstrates how the structural theory of causation clarifies relationships between probabilities and causes.

1.5.1 *The "Probability Raising" Trap*

The idea that causes raise the probability of their effects has been the engine behind most of PC explorations. It is a healthy idea, solidly ensconced in intuition. We say, for example, "reckless driving causes accidents" or "you will fail the course because of your laziness" (Suppes, 1970), knowing quite well that the antecedents merely tend to make the consequences more likely, not absolutely certain. One would expect, therefore, that probability raising should become the defining characteristic of the relationship between a cause (C) and its effect (E). Alas, though perfectly valid, this intuition cannot be expressed using the tools of probabilities; the relationship "raises the probability of" is counterfactual (or manipulative) in nature, and cannot, therefore, be captured in the language of probability theory.

The way philosophers tried to capture this relationship, using inequalities such as[15]

$$P(E|C) > P(E) \tag{1.13}$$

[15]Some authors write $P(E|C) > P(E|\neg C)$, which is equivalent to (1.13); the latter is easier to generalize to the non-binary case.

was misguided from the start – counterfactual "raising" cannot be reduced to evidential "raising," or "raising by conditioning." The correct inequality, according to the structural theory of Section 1.3, should read:

$$P(E|do(C)) > P(E) \tag{1.14}$$

where $do(C)$ stands for an external intervention that compels the truth of C. The conditional probability $P(E|C)$, as we know from Section1.3 represents a probability resulting from a passive observation of C, and rarely coincides with $P(E|do(C))$. Indeed, observing the barometer falling increases the probability of a storm coming, but does not "cause" the storm; if the act of manipulating the barometer were to change the probability of storms, the falling barometer would qualify as a cause of storms.

Reformulating the notion of "probability raising" within the calculus of *do*-operators resolves the difficulties that PC has encountered in the past half-century.[16] Two such difficulties are worth noting here, for they can be resolved by the analysis of Section 1.3.

1.5.2 *The mystery of "background context"*

Recognizing that the basic inequality $P(E|C) > P(E)$ may yield paradoxical results in the presence of confounding factors (e.g., the atmospheric pressure in the example above), philosophers have modified the inequality by conditioning on a background factor K, yielding the criterion: $P(E|C, K = k) > P(E|K = k)$ where K consists on a set of variables capable of creating spurious dependencies between the cause and the effect. However, the question of what variables should enter K led to speculations, controversies and fallacies.[17]

Cartwright (1983), for example, states that a factor F should enter into K if and only if F is *causally relevant* to the effect, that is, F tends to either promote or prevent E. Eells (1991) on the other hand dropped the "only if" part and insisted on the "if." The correct answer, as we know from our analysis of Section 1.3, is neither Cartwright's nor Eell's; K should merely satisfy the back-door criterion of Section 1.3.2, which may or may not include variables that are causally relevant to the effect E.

The background-context debate is symptomatic of the fundamental flaw of the probabilistic causality program; the program first misrepresented the causal relation $P(E|do(C))$ by a conditional probability surrogate $P(E|C)$, and then, to escape the wrath of spurious associations, attempted to patch-up the distortion by adding remedial conditionalizations, only to end up with a contested $P(E|C, K)$. The correct strategy should have been to define "probability raising"

[16] This paper focuses on "type causation" namely, the tendency of the cause to bring about the effect. Token causation, also known as "actual causation" (Pearl, 2000*a*, Chapter 10) requires heavier counterfactual machinery.

[17] Conditioning on *all* factors F preceding C (Good, 1961; Suppes, 1970) would lead to counter intuitive conclusions (Pearl, 2000*a*, p. 297).

directly in terms of the $do(x)$ operator (or counterfactual variables Y_x), which would have yielded general and coherent results with no need for remedies.[18]

1.5.3 *The epistemology of causal relevance and probability raising*

The introduction of a "causal relevance" relation into the definition of "cause" is of course circular, for it compromises the original goal of reducing causality to purely probabilistic relations. It gave rise however to an interesting epistemological problem whose aim is not reductive but interpretative: Given that humans store experience in the form of qualitative "causal relevance" relationships, (with variable X being "causally relevant" to Y whenever it can influence Y in some way), we ask whether this knowledge, together with a probability function P is sufficient for determining whether event $X = x$ is a cause of event $Y = y$ in the "probability raising" sense.[19]

The problem is interesting because it connects judgments of three different types: judgments about "causal relevance" (R), about probabilities (P), and about cause-effect relations (CE). There is little doubt that causal-relevance relationships form part of an agent epistemic state; such relationships are implied by people's understanding of mechanisms, and how mechanisms are put together in the world around them. It is also reasonable to assume that an agent's epistemic state contains some representation of a probability function P that summarizes facts, observations, and associations acquired by the agent, either directly or indirectly (say through hearsay, or reading scientific reports). Finally, people usually reach consensus judging whether a given event $X = x$ "causes" event $Y = y$, and generally agree with the "probability raising" maxim.

The epistemic question above amounts to asking whether the three types of judgments, R, P, and CE, are compatible with each other. Put differently, the question we may ask is whether CE judgments are compatible with the pair $< R, P >$ and the probability raising maxim given in (1.14). To answer such questions we must first determine whether the pair $< R, P >$ is sufficient for deriving inequalities of the type given in (1.14).

The structural theory of causation gives a definitive solution to this problem which reads as follows:
Given: A graph G on a set V of variables, such that there is a directed path from X to Y in V iff X is judged to be "causally relevant" to Y.
Also given: a probability measure $P(v)$ that is compatible with G.
Problem: Decide, for a given X and Y in V, whether the probability raising inequality (1.14) holds for $C : X = x$ and $E : Y = y$, namely whether the causal effect

---

[18] Lewis (1986) proposed indeed to treat probability raising in the context of his counterfactual theory. However, lacking structural semantics, PC advocates viewed Lewis's counterfactuals as resting on shaky formal foundation "for which we have only the beginnings of a semantics (via the device of measures over possible worlds)" (Cartwright, 1983, p. 34).

[19] This is my interpretation of Eell's (1991) epistemic consistency problem (Pearl, 2000*a*, p. 252).

$$CE = P(y|do(x)) - P(y) \tag{1.15}$$

is greater than zero, given G and P.

The solution follows immediately from the identification of causal effects in Markovian models, which permits the derivation of CE from G and P, for example, by the causal effect formula of equation (1.4).

The solution is less obvious when P is defined over a proper subset W of V, where $\{V - W\}$ represents the set of unmeasured variables. The problem then reduces to that of identifying CE in semi-Markovian models such as those addressed in Theorem 1. Fortunately, the completeness results of Tian and Pearl (2002) and Shpitser and Pearl (2006*b*) reduce this problem to algorithmic routine on the graph G and, furthermore, they provide a guarantee that, if the algorithm fails, then any algorithm would fail, namely the causal effect of x on y does not have a unique value, given R and P.

I venture to conjecture that *every* epistemic problem concerned with the relationship between causes and probabilities is now amenable to algorithmic solution, provided that one explicates formally what is assumed and what needs to be decided.

1.6 Comparison to the Potential-Outcomes Framework

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that outcome Y would obtain in experimental unit u, had treatment X been x" (Neyman, 1923; Rubin, 1974). Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In Section 1.3.3 we saw that this counterfactual entity has the natural interpretation as representing the solution for Y in a modified system of equations, where *unit* is interpreted a vector u of background factors that characterize an experimental unit. Each structural equation model thus carries a collection of assumptions about the behavior of hypothetical units, and these assumptions permit us to derive the counterfactual quantities of interest. In the potential-outcome framework, however, no equations are available for guidance and $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined; not a quantity that can be derived from some model. In this sense the structural interpretation of $Y_x(u)$ given in (1.7) provides the formal basis for the potential-outcome approach; the formation of the submodel M_x explicates mathematically how the hypothetical condition "had X been x" could be realized, and what the logical consequence are of such a condition.

1.6.1 *The "Black-Box" or "Missing-data" Paradigm*

The distinct characteristic of the potential-outcome approach is that, although investigators must think and communicate in terms of undefined, hypothetical quantities such as $Y_x(u)$, the analysis itself is conducted almost entirely within the axiomatic framework of probability theory. This is accomplished, by postulating a "super" probability function on both hypothetical and real events. If U is treated as a random variable then the value of the counterfactual $Y_x(u)$

becomes a random variable as well, denoted as Y_x. The potential-outcome analysis proceeds by treating the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects (written $P(y|do(x))$ in the structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y_x = y)$. The new hypothetical entities Y_x are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence.

Naturally, these hypothetical entities are not entirely whimsy. They are assumed to be connected to observed variables via consistency constraints (Robins, 1986) such as

$$X = x \implies Y_x = Y, \tag{1.16}$$

which states that, for every u, if the actual value of X turns out to be x, then the value that Y would take on if 'X *were* x' is equal to the actual value of Y. For example, a person who chose treatment x and recovered, would also have recovered if given treatment x by design. Whether additional constraints should tie the observables to the unobservables is not a question that can be answered in the potential-outcome framework, which lacks an underlying model.

The main conceptual difference between the two approaches is that, whereas the structural approach views the intervention $do(x)$ as an operation that changes the distribution but keeps the variables the same, the potential-outcome approach views the variable Y under $do(x)$ to be a different variable, Y_x, loosely connected to Y through relations such as (1.16), but remaining unobserved whenever $X \neq x$. The problem of inferring probabilistic properties of Y_x, then becomes one of "missing-data" for which estimation techniques have been developed in the statistical literature.

Pearl (2000*a*, Chapter 7) shows, using the structural interpretation of $Y_x(u)$, that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (1.16) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \text{ for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{1.17}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \text{ for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{1.18}$$

Equation (1.17) ensures that the interventions $do(Y = y)$ results in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that may be applied to variables other than Y. Equation (1.18) generalizes (1.16) to cases where Z is held fixed, at z.

1.6.2 *Problem Formulation and the Demystification of "Ignorability"*

The main drawback of this black-box approach surfaces in problem formulation, namely, the phase where a researcher begins to articulate the "science" or "causal

assumptions" behind the problem at hand. Such knowledge, as we have seen in Section 1.1, must be articulated at the onset of every problem in causal analysis – causal conclusions are only as valid as the causal assumptions upon which they rest.

To communicate scientific knowledge, the potential-outcome analyst must express assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For instance, in our example of Figure 1.2(a), to communicate the understanding that the Z is randomized (hence independent of U_X and U_Y), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{Y_{z_1}, Y_{z_2}, \ldots, Y_{z_k}\}$.[20] To further formulate the understanding that Z does not affect Y directly, except through X, the analyst would write a, so called, "exclusion restriction": $Y_{xz} = Y_x$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest. For example, if one can plausibly assume that, in Fig. 1.3, a set Z of covariates satisfies the conditional independence

$$Y_x \perp\!\!\!\perp X | Z \tag{1.19}$$

(an assumption termed "conditional ignorability" by Rosenbaum and Rubin (1983),) then the causal effect $P(y|do(x)) = P^*(Y_x = y)$ can readily be evaluated to yield

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x,z)P(z) \quad \text{(using (1.19))} \\ &= \sum_z P^*(Y = y|x,z)P(z) \quad \text{(using (1.16))} \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{1.20}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the standard covariate-adjustment formula of equation (1.5).

We see that the assumption of conditional ignorability (1.19) qualifies Z as a sufficient covariate for adjustment; it is entailed indeed by the "back-door" criterion of Section 1.3.2, which qualifies such covariates by tracing paths in the causal diagram.

The derivation above may explain why the potential-outcome approach appeals to mathematical statisticians; instead of constructing new vocabulary (e.g., arrows), new operators ($do(x)$) and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of

[20] The notation $Y \perp\!\!\!\perp X|Z$ stands for the conditional independence relationship $P(Y = y, X = x|Z = z) = P(Y = y|Z = z)P(X = x|Z = z)$ (Dawid, 1979).

probability calculus. Save for an occasional application of rule (1.18) or (1.16)), the analyst may forget that Y_x stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

However, this mathematical orthodoxy exacts a very high cost: all background knowledge pertaining to a given problem must first be translated into the language of counterfactuals (e.g., ignorability conditions) before analysis can commence. This translation may in fact be the hardest part of the problem. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (1.19), the key to the derivation of (1.20), holds in any familiar situation, say in the experimental setup of Figure 1.2(a). This assumption reads: "the value that Y would obtain had X been x, is independent of X, given Z". Even the most experienced potential-outcome expert would be unable to discern whether any subset Z of covariates in Figure 1.3 would satisfy this conditional independence condition.[21] Likewise, to derive equation (1.6) in the language of potential-outcome (see Pearl (2000*a*, p. 223)), one would need to convey the structure of the chain $X \rightarrow W_3 \rightarrow Y$ using the cryptic expression: $W_{3_x} \perp\!\!\!\perp \{Y_{w_3}, X\}$, read: "the value that W_3 would obtain had X been x is independent of the value that Y would obtain had W_3 been w_3 jointly with the value of X". Such assumptions are cast in a language so far removed from ordinary understanding of scientific theories that, for all practical purposes, they cannot be comprehended or ascertained by ordinary mortals. As a result, researchers in the graph-less potential-outcome camp rarely use "conditional ignorability" (1.19) to guide the choice of covariates; they view this condition as a hoped-for miracle of nature rather than a target to be achieved by reasoned design.[22]

Replacing "ignorability" with a simple condition (i.e., back-door) in a graphical model permits researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are selected, and what experiments we can do to test, at least partially, if we have the knowledge needed for covariate selection.

Aside from offering no guidance in covariate selection, formulating a problem in the potential-outcome language encounters three additional hurdles. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are *redundant*, or whether those judgments are *self-consistent*. The need to express, defend, and manage formidable counterfactual relationships of this type explain the slow

[21] Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z|Y$ holds in Figure 1.2(a).

[22] The opaqueness of counterfactual independencies explains why many researchers within the potential-outcome camp are unaware of the fact that adding a covariate to the analysis (e.g., Z_3 in Figure 1.3) may actually *increase* confounding bias. Paul Rosenbaum, for example, writes: "there is no reason to avoid adjustment for a variable describing subjects before treatment" (Rosenbaum, 2002, p. 76). Rubin (2009) goes as far as stating that refraining from conditioning on an available measurement is "nonscientific ad hockery" for it goes against the tenets of Bayesian philosophy (see (Pearl, 2009*bc*) for a discussion of this fallacy).

acceptance of causal analysis among health scientists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential-outcome alternatives advocated in (Angrist *et al.*, 1996; Holland, 1988; Sobel, 1998).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist *et al.*, 1996), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian and Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2000*a*). Pearl (2000*a*, p. 232) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into counterfactual notation, performing the mathematics in the algebraic language of counterfactuals (using (1.16), (1.17), and (1.18)) and, finally, interpreting the result in plain causal language. The mediation problem of Section 1.4 illustrates such symbiosis.

In comparison, when the mediation problem is approached from an orthodox potential-outcome viewpoint, void of the structural guidance of equation (1.7), pardoxical results ensue (Rubin, 2004). For example, the direct effect is definable only in units absent of indirect effects. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This leaves us mostly with odd families, absent of grandfathers or fathers. In linear systems, to take a sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. Such paradoxical conclusions underscore the wisdom, if not necessity of a symbiotic analysis, in which the counterfactual notation $Y_x(u)$ is governed by the structural semantics of the SCM.

1.7 Conclusions

Theories of causation require two ingredients that are absent from probabilistic or logical theories; a science-friendly language for articulating causal knowledge, and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomenon. This paper introduces a general theory of causation, based on nonparametric structural equations models, that supplements statistical methods with the needed ingredients. The algebraic component of the theory coincides with the potential-outcome framework, and its graphical component embraces Wright's method of path diagrams (in its nonparametric version). When unified and synthesized, the two components offer empirical investigators a powerful and comprehensive methodology for causal inference. and a general framework for viewing other, less general approaches to causation, including probabilistic causation (Section 1.5) and the potential-outcome model (1.6).

Acknowledgments

Portions of this paper are based on my book *Causality* (Pearl, 2000, 2nd edition forthcoming 2009), and have benefited appreciably from conversations with Chris Hitchcock. This research was supported in parts by an ONR grant #N000-14-09-1-0665.

REFERENCES

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, **91**(434), 444–472.

Arah, O.A. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology*, **4**, doi:10.1186/1742–7622–5–5. Online at http://www.ete-online.com/content/5/1/5.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, Edinburgh, UK, pp. 357–363. Morgan-Kaufmann Publishers.

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

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

Balke, A. and Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence 11* (ed. P. Besnard and S. Hanks), pp. 11–18. Morgan Kaufmann, San Francisco.

Brent, R. and Lok, L. (2005). A fishing buddy for hypothesis generators. *Science*, **308**(5721), 523–529.

Cartwright, N. (1983). *How the Laws of Physics Lie.* Clarendon Press, Oxford.

Chalak, K. and White, H. (2006, July). An extended class of instrumental variables for the estimation of causal effects. Technical Report Discussion Paper, UCSD, Department of Economics.

Cole, S.R. and Hernán, M.A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, **31**(1), 163–165.

Collins, J., Hall, N., and Paul, L.A. (eds.) (2004). *Causation and Counterfactuals.* MIT Press, Cambridge, MA.

Cox, D.R. (1958). *The Planning of Experiments.* John Wiley and Sons, NY.

Cox, D.R. and Wermuth, N. (2004). Causality: A statistical view. *International Statistical Review*, **72**(3), 285–305.

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

Dawid, A.P. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association*, **95**(450), 407–

448.

Dawid, A.P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–189.

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

Eells, E. (1991). *Probabilistic Causality.* Cambridge University Press, Cambridge, MA.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)*, **69**(2), 199–215.

Glymour, M.M. and Greenland, S. (2008). Causal diagrams. In *Modern Epidemiology* (3rd edn) (ed. K. Rothman, S. Greenland, and T. Lash). Lippincott Williams & Wilkins, Philadelphia, PA.

Good, I.J. (1961). A causal calculus (I). *British Journal for the Philosophy of Science*, **11**, 305–318.

Greenland, S. and Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, **31**, 1030–1037.

Greenland, S., Pearl, J., and Robins, J.M (1999). Causal diagrams for epidemiologic research. *Epidemiology*, **10**(1), 37–48.

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

Heckman, J.J. (2008). Econometric causality. *International Statistical Review*, **76**(1), 1–27.

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

Hitchcock, C.R. (2003). Probabilistic causation. In *Stanford Encyclopedia of Philosophy (Winter 2003 Edition)* (ed. E. Zalta). URL = <http://plato.stanford.edu/entries/causation-probabilistic/>.

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

Lauritzen, S.L. (1996). *Graphical Models.* Clarendon Press, Oxford.

Lauritzen, S.L. (2001). Causal inference from graphical models. In *Complex Stochastic Systems* (ed. D. Cox and C. Kluppelberg), pp. 63–107. Chapman and Hall/CRC Press, Boca Raton, FL.

Lewis, D. (1986). *Philosophical Papers*, Volume II. Oxford University Press, New York.

Lindley, D.V. (2002). Seeing and doing: The concept of causation. *International Statistical Review*, **70**, 191–214.

Meek, C. and Glymour, C.N. (1994). Conditioning and intervening. *British Journal of Philosophy Science*, **45**, 1001–1021.

Morgan, S.L. and Winship, C. (2007). *Counterfactuals and Causal Inference:*

Methods and Principles for Social Research (Analytical Methods for Social Research). Cambridge University Press, New York, NY.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**(4), 465–480.

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

Pearl, J. (1993*a*). Comment: Graphical models, causality, and intervention. *Statistical Science*, **8**(3), 266–269.

Pearl, J. (1993*b*). Mediating instrumental variables. Technical Report Technical Report R-210, Computer Science Department, UCLA.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**(4), 669–710.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, **27**(2), 226–284.

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

Pearl, J. (2000*b*). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association*, **95**(450), 428–431.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann, San Francisco, CA.

Pearl, J. (2003, December). Statistics and causal inference: A review. *Test Journal*, **12**(2), 281–345.

Pearl, J. (2005). Direct and indirect effects. In *Proceedings of the American Statistical Association, Joint Statistical Meetings*, pp. 1572–1581. MIRA Digital Publishing, Minn., MN.

Pearl, J. (2009*a*). *Causality: Models, Reasoning, and Inference* (Second edn). Cambridge University Press, New York. Forthcoming.

Pearl, J. (2009*b*). Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine*, **28**, 1415–1416. http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf.

Pearl, J. (2009*c*). Myth, confusion, and science in causal analysis. Technical Report R-348, University of California, Los Angeles, CA. http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf.

Pearl, J. and Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (ed. P. Besnard and S. Hanks), pp. 444–453. Morgan Kaufmann, San Francisco.

Petersen, M.L., Sinisi, S.E., and van der Laan, M.J. (2006). Estimation of direct causal effects. *Epidemiology*, **17**(3), 276–284.

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

Robins, J.M. (1999). Testing and estimation of directed effects by reparameterizing directed acyclic with structural nested models. In *Computation, Causation, and Discovery* (ed. C. Glymour and G. Cooper), pp. 349–405. AAAI/MIT Press, Cambridge, MA.

Robins, J.M. and Greenland, S. (1991). Estimability and estimation of expected years of life lost due to a hazardous exposure. *Statistics in Medicine*, **10**, 79–93.

Rosenbaum, P.R. (2002). *Observational Studies* (Second edn). Springer-Verlag, New York.

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rothman, K.J. (1976). Causes. *American Journal of Epidemiology*, **104**, 587–592.

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

Rubin, D.B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, **31**, 161–170.

Rubin, D.B. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine*, **28**, 1420–1423.

Shpitser, I. and Pearl, J (2006*a*). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (ed. R. Dechter and T. Richardson), pp. 437–444. AUAI Press, Corvallis, OR.

Shpitser, I. and Pearl, J. (2006*b*). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pp. 1219–1226. AAAI Press, Menlo Park, CA.

Shpitser, I. and Pearl, J (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359. AUAI Press, Vancouver, BC, Canada. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shrier, I. (2009). Letter to the editor: Propensity scores. *Statistics in Medicine*, **28**, 1317–1318. See also Pearl 2009 <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

Simon, H.A. and Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science*, **33**, 323–340.

Sobel, M.E. (1998). Causal inference in statistical models of the process of socioeconomic achievement. *Sociological Methods & Research*, **27**(2), 318–348.

Spirtes, P., Glymour, C.N., and Scheines, R. (2000). *Causation, Prediction, and Search* (2nd edn). MIT Press, Cambridge, MA.

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

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

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

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

Williamson, J. (2010, Forthcoming). Probabilistic theories of causality. In *The Oxford Handbook of Causation* (ed. H. Beebee, C. Hitchcock, and P. Peter). Oxford University Press, New York.

Woodward, J. (2003). *Making Things Happen.* Oxford University Press, New York, NY.

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