In Joint Statistical Meetings Proceedings,
Alexandria, VA: American Statistical Association,
2515-2529, 2013.

TECHNICAL REPORT
R-416
September 2013

The Mathematics of Causal Inference

Judea Pearl*

Abstract
This paper reviews concepts, principles and tools that have led to a coherent mathematical theory of causation based on structural models. The theory provides solutions to a number of problems in causal inference, including questions of confounding control, policy analysis, mediation, missing data and the integration of data from diverse studies.

Key Words: Causal Inference, confounding, mediation, missing data, meta analysis

1. Introduction

Recent advances in causal inference owe their development to two methodological principles. First, a commitment to understanding what reality must be like for a statistical routine to succeed and, second, a commitment to represent reality in terms of data-generating models, rather than distributions of observed variables.

Encoded as nonparametric structural equations, these models have led to a fruitful symbiosis between graphs and counterfactuals that has unified the potential outcome framework of Neyman, Rubin and Robins with the econometric tradition of Haavelmo, Marschak and Heckman. In this symbiosis, counterfactuals emerge as natural byproducts of structural equations and serve to formally articulate research questions of interest. Graphical models, on the other hand, are used to encode scientific assumptions in a qualitative (i.e., nonparametric) and transparent language, and to identify the testable implications of these assumptions.

In Section 2 we define Structural Causal Models (SCM) and state the two fundamental laws of causal inference: (1) how counterfactuals and probabilities of counterfactuals are deduced from a given SCM and (2) how features of the observed data are shaped by the graphical structure of a SCM.

Section 3 defines the challenge of *identifying* causal parameters and presents a complete solution to the problem of nonparametric identification of causal effects. Given data from observational studies and qualitative assumptions in the form of a graph with measured and unmeasured variables, we need to decide algorithmically whether the assumptions are sufficient for identifying causal effects of interest, what covariates should be measured and what the statistical estimand is of the identified effect.

Section 4 summarizes mathematical results concerning nonparametric *mediation*, which aims to estimate the extent to which a given effect is mediated by various pathways or mechanisms. A simple set of conditions will be presented for estimating *natural* direct and indirect effects in nonparametric models.

Section 5 deals with the problem of "generalizability" or "external validity": under what conditions can we take experimental results from one or several populations and apply them to another population which is potentially different from the rest. A complete solution to this problem will be presented in the form of an algorithm which decides whether a specific causal effect is transportable and, if the answer is affirmative, what measurements need be taken in the various populations and how they ought to be combined.

*UCLA Computer Science Department, 4532 Boelter Hall, Los Angeles, CA 90095-1596

Finally, Section 6 describes recent work on missing data and shows that, by viewing missing data as a causal inference task, the space of problems can be partitioned into two algorithmically recognized categories: those that permit consistent recovery from missingness and those that do not.

To facilitate clarity and accessibility the major mathematical results will be highlighted in the form of four "Summary Results," and will be framed in boxes.

2. Counterfactuals and the Structural Causal Model (SCM)

At the center of the structural theory of causation lies a "structural model," M, consisting of two sets of variables, U and V, and a set F of functions that determine, or simulate how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which variable V_i is *assigned* the value $v_i = f_i(v, u)$ in response to the current values, v and u, of all variables in V and U. Formally, the triplet $< U, V, F >$ defines a structural causal model (SCM), and the diagram, that captures the relationships among the variables is called the *causal graph* G (of M). The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. Letting M_x stand for a modified version of M, with the equation(s) of set X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{1}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "modified" submodel M_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models (see also Pearl, 2009b, Chapter 7).

Since the distribution $P(u)$ induces a well defined probability on the counterfactual event $Y_x = y$, it also defines a joint distribution on all Boolean combinations of such events, for instance '$Y_x = y$ AND $Z_{x'} = z$,' which may appear contradictory, if $x \neq x'$. For example, to answer retrospective 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.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any propositional evidence, can be computed by the 3-step process (illustrated in Pearl, 2009b, p. 207);

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations determining the variables in set X by $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

2.1 The two principles of causal inference

Before describing specific applications of the structural theory, it will be useful to summarize its implications in the form of two "principles," from which all other results follow.

Principle 1: "The law of structural counterfactuals."

Principle 2: "The law of structural independence."

The first principle is described in Eq. (1) and instructs us how to compute counterfactuals and probabilities of counterfactuals from a structural model. This, together with principle 2 will allow us (Section 3) to determine what assumptions one must make about reality in order to infer probabilities of counterfactuals from either experimental or passive observations.

Principle 2, defines how structural features of the model entail dependencies in the data. Remarkably, regardless of the functional form of the equations in the model and regardless of the distribution of the exogenous variables U, if the latters are mutually independent and the model is recursive, the distribution $P(v)$ of the endogenous variables must obey certain conditional independence relations, stated roughly as follows: whenever sets X and Y are "separated" by a set Z in the graph, X is independent of Y given Z in the probability.

This "separation" condition, called d-separation (Pearl, 2000a, pp. 16–18) constitutes the link between the causal assumptions encoded in the causal graph (in the form of missing arrows) and the observed data.

Definition 1. *(d-separation)*
A set S of nodes is said to block a path p if either

1. *p contains at least one arrow-emitting node that is in S, or*
2. *p contains at least one collision node that is outside S and has no descendant in S.*

If S blocks all *paths from set X to set Y, it is said to "d-separate X and Y," and then, variables X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.*[1]

D-separation implies conditional independencies for every distribution $P(v)$ that is compatible with the causal assumptions embedded in the diagram. To illustrate, the diagram in Fig. 1(a) implies $Z_1 \perp\!\!\!\perp Y|(X, Z_3, W_2)$, because the conditioning set $S = \{X, Z_3, W_2\}$ blocks all paths beween Z_1 and Y. The set $S = \{X, Z_3, W_3\}$ however leaves the path (Z_1, Z_3, Z_2, W_2, Y) unblocked (by virtue of the collider at Z_3) and, so, the independence $Z_1 \perp\!\!\!\perp Y|(X, Z_3, W_3)$ is not implied by the diagram.

3. Intervention, Identification, and Causal Calculus

A central question in causal analysis is that of predicting the results of interventions, such as those resulting from treatments or social programs, which we denote by the symbol $do(x)$ and define using the counterfactual Y_x as[2]

$$P(y|do(x)) \triangleq P(Y_x = y) \tag{2}$$

[1] By a "path" we mean a consecutive edges in the graph regardless of direction. See Pearl (2009b, p. 335) for a gentle introduction to d-separation and its proof. In linear models, the independencies implied by d-separation are valid for non-recursive models as well.

[2] An alternative definition of $do(x)$, invoking population averages only, is given in (Pearl, 2009b, p. 24).

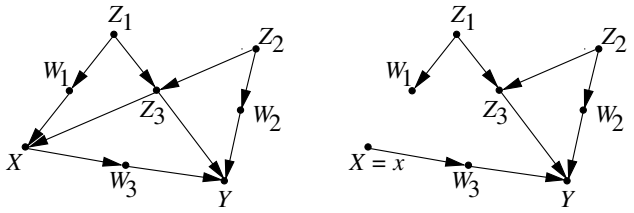

Figure 1: (a) Graphical model illustrating d-separation and the back-door criterion. U terms are not shown explicitly. (b) Illustrating the intervention $do(X = x)$.

Figure 2(b) illustrates the submodel M_x created by the atomic intervention $do(x)$; it sets the value of X to x and thus removes the influence of W_1 and Z_3 on X. We similarly define the result of *conditional interventions* by

$$P(y|do(x), z) \stackrel{\Delta}{=} P(y, z|do(x))/P(z|do(x) = P(Y_x = y|Z_x = z) \tag{3}$$

$P(y|do(x), z)$ captures the z-specific effect of X on Y, that is, the effect of setting X to x among those units only for which $Z = z$.

A second important question concerns *identification* in partially specified models: Given a set A of qualitative causal assumptions, as embodied in the structure of the causal graph, can the controlled (post-intervention) distribution, $P(y|do(x))$, be estimated from the available data which is governed by the pre-intervention distribution $P(z, x, y)$? In linear parametric settings, the question of identification reduces to asking whether some model parameter, β, has a unique solution in terms of the parameters of P (say the population covariance matrix). In the nonparametric formulation, the notion of "has a unique solution" does not directly apply since quantities such as $Q = P(y|do(x))$ have no parametric signature and are defined procedurally by a symbolic operation on the causal model M, as in Fig. 1(b). The following definition captures the requirement that Q be estimable from the data:

Definition 2. *(Identifiability) (Pearl, 2000a, p. 77)*
A causal query Q is identifiable from data compatible with a causal graph G, if for any two (fully specified) models M_1 and M_2 that satisfy the assumptions in G, we have

$$P_1(v) = P_2(v) \Rightarrow Q(M_1) = Q(M_2) \tag{4}$$

In words, equality in the probabilities $P_1(v)$ and $P_2(v)$ induced by models M_1 and M_2, respectively, entails equality in the answers that these two models give to query Q. When this happens, Q depends on P only, and should therefore be expressible in terms of the parameters of P.

When a query Q is given in the form of a do-expression, for example $Q = P(y|do(x), z)$, its identifiability can be decided systematically using an algebraic procedure known as the do-calculus (Pearl, 1995). It consists of three inference rules that permit us to equate interventional and observational distributions whenever certain d-separation conditions hold in the causal diagram G.

3.1 The Rules of *do*-calculus

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise,

we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$.

The following three rules are valid for every interventional distribution compatible with G.

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{5}$$

Rule 2 (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{6}$$

Rule 3 (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}, \tag{7}$$

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

To establish identifiability of a causal query Q, one needs to repeatedly apply the rules of do-calculus to Q, until an expression is obtained which no longer contains a do-operator[3]; this renders it estimable from nonexperimental data. The do-calculus was proven to be complete for queries in the form $Q = P(y|do(x), z)$ (Huang and Valtorta, 2006; Shpitser and Pearl, 2006), which means that if Q cannot be reduced to probabilities of observables by repeated application of these three rules, such a reduction does not exist, i.e., the query is not estimable from observational studies without strengthening the assumptions.

Covariate Selection: The back-door criterion

Consider an observational study where we wish to find the effect of treatment (X) on outcome (Y), and assume that the factors deemed relevant to the problem are structured as in Fig. 1(a); some are affecting the outcome, some are affecting the treatment, and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or lifestyle, while others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment such 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," "admissible set" or a set "appropriate for adjustment" (see Greenland et al. 1999; Pearl 2000b, 2009a).

The following criterion, named "back-door" (Pearl, 1993) provides a graphical method of selecting such a set of factors for adjustment.

Definition 3. *(admissible sets—the back-door criterion)*
A set S is admissible (or "sufficient") for estimating the causal effect of X on Y 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.*

Based on this criterion we see, for example that, in Fig. 1, the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$ are each 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 it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

[3]Such derivations are illustrated in graphical details in (Pearl, 2009b, p. 87).

The intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from X to Y, while the paths directed along the arrows from X to Y carry causative associations. Blocking the former paths (by conditioning on S) ensures that the measured association between X and Y is purely causal, namely, it correctly represents the target quantity: the causal effect of X on Y. Conditions for relaxing restriction 1 are given in (Pearl, 2009b, p. 338; Shpitser et al., 2010).

The implication of finding a sufficient set, S, is that stratifying on S is guaranteed to remove all confounding bias relative to the causal effect of X on Y. In other words, it renders the causal effect of X on Y identifiable, via the *adjustment formula*[4]

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \tag{8}$$

Since all factors on the right-hand side of the equation are estimable (e.g., by regression) from pre-interventional data, the causal effect can likewise be estimated from such data without bias. Moreover, the back-door criterion implies the independence $X \perp\!\!\!\perp Y_x|S$, also known as "conditional ignorability" (Rosenbaum and Rubin, 1983) and, provides therefore the scientific basis for most inferences in the potential outcome framework.

The back-door criterion allows us to write Eq. (8) by inspection, after selecting a sufficient set, S, from the diagram. 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-response framework. The criterion also enables the analyst to search for an optimal set of covariates—namely, a set, S, that minimizes measurement cost or sampling variability (Tian et al., 1998).

Summary Result 1. *(Identification of Interventional Expressions) Given a causal graph G containing both measured and unmeasured variables, the consistent estimability of any expression of the form*

$$Q = P(y_1, y_2, \ldots, y_m|do(x_1, x_2, \ldots, x_n), z_1, z_2, \ldots, z_k)$$

can be decided in polynomial time. If Q is estimable, then its estimand can be derived in polynomial time. Furthermore, the algorithm is complete.

The results stated in Summary Result 1 were developed in several stages over the past 20 years (Pearl, 1993, 1995; Tian and Pearl, 2002; Shpitser and Pearl, 2006). Bareinboim and Pearl (2012a) extended the identifiability of Q to combinations of observational and experimental studies.

4. Mediation Analysis

The nonparametric structural model for a typical mediation problem takes the form:

$$t = f_{_T}(u_{_T}) \qquad m = f_{_M}(t, u_{_M}) \qquad y = f_{_Y}(t, m, u_{_Y}) \tag{9}$$

where T (treatment), M (mediator), and Y (outcome) are discrete or continuous random variables, $f_{_T}, f_{_M}$, and $f_{_Y}$ are arbitrary functions, and $U_{_T}, U_{_M}, U_{_Y}$ represent respectively omitted factors that influence T, M, and Y. The triplet $U = (U_{_T}, U_{_M}, U_{_Y})$ is a random vector that accounts for all variations between individuals. It is sometimes called "unit," for it offers a complete characterization of a subject's behavior as reflected in T, M, and Y.

[4]Summations should be replaced by integration when applied to continuous variables, as in (Imai et al., 2010).

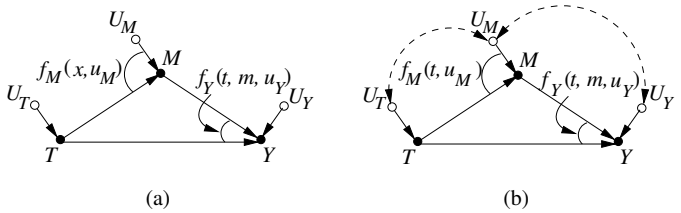

Figure 2: (a) The basic nonparametric mediation model, with no confounding. (b) A confounded mediation model in which dependence exists between U_M and (U_T, U_Y).

The distribution of U, denoted $P(U = u)$, uniquely determines the distribution $P(t, m, y)$ of the observed variables through the three functions in Eq. (9).

In Fig. 2(a) the omitted factors are assumed to be arbitrarily distributed but mutually independent, written $U_T \perp\!\!\!\perp U_M \perp\!\!\!\perp U_Y$. In Fig. 2(b) the dashed arcs connecting U_T and U_M (as well as U_M and U_T) encode the understanding that the factors in question may be dependent.

4.1 Natural direct and indirect effects

Using the structural model of Eq. (9), four types of effects can be defined for the transition from $T = 0$ to $T = 1$:[5]

(a) Total Effect –

$$\begin{aligned} TE &= E\{f_Y[1, f_M(1, u_M), u_Y] - f_Y[0, f_M(0, u_M), u_Y]\} \\ &= E[Y_1 - Y_0] \\ &= E[Y|do(T = 1)] - E[Y|do(T = 0)] \end{aligned} \tag{10}$$

TE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T as dictated by the function f_M.

(b) Controlled Direct Effect –

$$\begin{aligned} CDE(m) &= E\{f_Y[1, M = m, u_Y] - f_Y[0, M = m, u_Y]\} \\ &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y|do(T = 1, M = m)] - E[Y|do(T = 0, M = m)] \end{aligned} \tag{11}$$

CDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to a specified level $M = m$ uniformly over the entire population.

(c) Natural Direct Effect –

$$\begin{aligned} NDE &= E\{f_Y[1, f_M(0, u_M), u_T] - f_Y[0, f_M(0, u_M), u_T]\} \\ &= E[Y_{1,M_0} - Y_{0,M_0}] \end{aligned} \tag{12}$$

NDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it *would have attained* (for each individual) prior to the change, i.e., under $T = 0$.

(d) Natural Indirect Effect –

$$\begin{aligned} NIE &= E\{f_Y[0, f_M(1, u_M), u_Y] - f_Y[0, f_M(0, u_M), u_Y]\} \\ &= E[Y_{0,M_1} - Y_{0,M_0}] \end{aligned} \tag{13}$$

[5]Generalizations to arbitrary reference point, say from $T = t$ to $T = t'$, are straightforward. These definitions apply at the population levels; the unit-level effects are given by the expressions under the expectation. All expectations are taken over the factors U_M and U_Y.

NIE measures the expected increase in Y when the treatment is held constant, at $T = 0$, and M changes to whatever value it would have attained (for each individual) under $T = 1$. It captures, therefore, the portion of the effect which can be explained by mediation alone, while disabling the capacity of Y responds to X.

We note that, in general, the total effect can be decomposed as

$$TE = NDE - NIE_r \tag{14}$$

where NIE_r stands for the natural indirect effect under the reverse transition, from $T = 1$ to $T = 0$. This implies that NIE is identifiable whenever NDE and TE are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula, $TE = NDE + NIE$.

We further note that TE and $CDE(m)$ are do-expressions and can, therefore be estimated from experimental data; not so NDE and NIE. Since Summary Result 1 assures us that the identifiability of any do-expression can be determined by an effective algorithm, we will regard the identifiability of TE and $CDE(m)$ as solved problems, and will focus our attention on NDE and NIE.

4.2 Sufficient conditions for identifying natural effects

The following is a set of assumptions or conditions, marked A-1 to A-4, that are sufficient for identifying both direct and indirect natural effects. Each condition is communicated by a verbal description followed by its formal expression. The full set of assumptions is then followed by its graphical representation.

Assumption set A (Pearl, 2001)

There exists a set W of measured covariates such that:

A-1 No member of W is affected by treatment.

A-2 W deconfounds the mediator-outcome relationship (holding T constant) i.e.,

$$[M_t \perp\!\!\!\perp Y_{t',m} \mid W]$$

A-3 The W-specific effect of the treatment on the mediator is identifiable by some means.

$$[P(m \mid do(t), w) \text{ is identifiable}]$$

A-4 The W-specific joint effect of {treatment+mediator} on the outcome is identifiable by some means.

$$[P(y \mid do(t, m), w) \text{ is identifiable}]$$

Graphical version of assumption set A

There exists a set W of measured covariates such that:

A_G-1 No member of W is a descendant of T.

A_G-2 W blocks all back-door paths from M to Y (not traversing $X \to M$ and $X \to Y$).

A_G-3 The W-specific effect of T on M is identifiable (using Summary Result 1, and possibly using experiments or auxiliary variables).

A_G-4 The W-specific joint effect of $\{T, M\}$ on Y is identifiable (using Summary Result 1, and possibly using experiments or auxiliary variables).

Summary Result 2. *(Identification of natural effects)*
When conditions A-1 and A-2 hold, the natural direct effect is experimentally identifiable and is given by

$$NDE = \sum_m \sum_w [E(Y|do(T=1, M=m)), W=w) - E(Y|do(T=0, M=m), W=w)] \\ P(M=m|do(T=0), W=w)P(W=w) \quad (15)$$

The identifiability of the do-expressions in (15) is guaranteed by conditions A-3 and A-4, and can be determined by Summary Result 1.

In the non-confounding case (Fig. 2(a)) NDE reduces to the mediation formula*:*

$$NDE = \sum_m [E(Y \mid T=1, M=m) - E(Y \mid T=0, M=m)]P(M=m \mid T=0). \quad (16)$$

Corollary 1. *If conditions A-1 and A-2 are satisfied by a set W that also deconfounds the relationships in A-3 and A-4, then the do-expressions in (15) are reducible to conditional expectations, and the natural direct effect becomes:*[6]

$$NDE = \sum_m \sum_w [E(Y|T=1, M=m, W=w) - E(Y|T=0, M=m, W=w)] \\ P(M=m|T=0, W=w)P(W=w) \quad (17)$$

It is interesting to compare assumptions A-1 to A-4 to those often cited in the literature, which are based on "sequential ignorability" (Imai et al., 2010), the dominant inferential tool in the potential outcome framework.

Assumption set B (Sequential ignorability)

There exists a set W of measured covariates such that:

B-1 W and T deconfound the mediator-outcome relationship.

$$[Y_{t',m} \perp\!\!\!\perp M_t \mid T, W]$$

B-2 W deconfounds the treatment-{mediator, outcome} relationship.

$$[T \perp\!\!\!\perp (Y_{t',m}, M_t) \mid W]$$

Assumption set A differs from assumption set B on two main provisions. First, A-3 and A-4 permit the identification of these causal effects by any means, while B-1 and B-2 insist that identification be accomplished by adjustment for W only. Second, whereas A-3 and A-4 auxiliary covariates to be invoked in the identification of the causal effects needed, B requires that the same set W satisfy all conditions simultaneously. Due to these two provisions, assumption set A significantly broadens the class of problems in which the natural effects are identifiable (Pearl, 2013). Shpitser (2013) further provides complete algorithms for identifying natural direct and indirect effects and extends these results to path-specific effects with multiple treatments and multiple outcomes.

[6]Equation (17) is identical to the one derived by Imai et al. (2010) using sequential ignorability (i.e., assumptions B-1 and B-2) and subsequently re-derived by a number of other authors (Wang and Sobel, 2013).

5. External Validity and Transportability

In applications requiring identification, the role of the *do*-calculus is to remove the *do*-operator from the query expression. We now discuss a totally different application, to decide if experimental findings from environment π can be transported to a new, potentially different environment, π^*, where only passive observations can be performed. This problem, labeled "transportability" in (Pearl and Bareinboim, 2011) is at the heart of every scientific investigation since, invariably, experiments performed in one environment (or population) are intended to be used elsewhere, where conditions may differ.

To formalize problems of this sort, a graphical representation called "selection diagrams" was devised (Fig. 3) which encodes knowledge about differences and commonalities between populations. A selection diagram is a causal diagram annotated with new variables, called S-nodes, which point to the mechanisms where discrepancies between the two populations are suspected to take place. The task of deciding if transportability

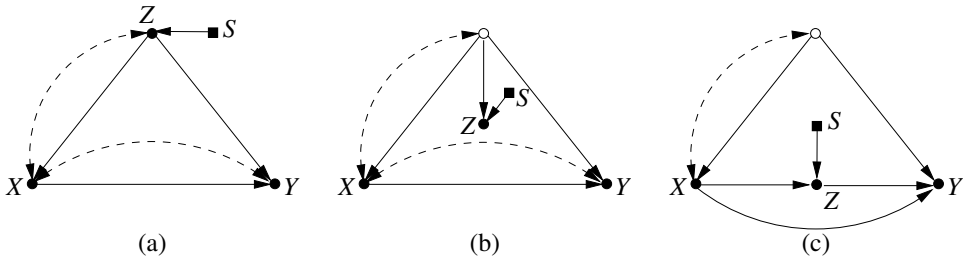

Figure 3: Selection diagrams depicting differences in populations. In (a) the two populations differ in age distributions. In (b) the populations differs in how reading skills (Z) depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c) the populations differ in how Z depends on X. Dashed arcs (e.g., $X \dashleftarrow\!\!\dashrightarrow Y$) represent the presence of latent variables affecting both X and Y.

is feasible now reduces to a syntactic problem of separating (using the *do*-calculus) the *do*-operator from a the S-variables in the query expression $P(y|do(x), z, s)$.

Theorem 1. *(Pearl and Bareinboim, 2011) Let D be the selection diagram characterizing two populations, π and π^*, and S a set of selection variables in D. The relation $R = P^*(y|do(x), z)$ is transportable from π and π^* if and only if the expression $P(y|do(x), z, s)$ is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.*

While Theorem 1 does not specify the sequence of rules leading to the needed reduction (if such exists), a complete and effective graphical procedure was devised by Bareinboim and Pearl (2012b), which also synthesizes a *transport formula* whenever possible. Each transport formula determines what information need to be extracted from the experimental and observational studies and how they ought to be combined to yield an unbiased estimate of the relation $R = P(y|do(x), s)$ in the target population π^*. For example, the transport formulas induced by the three models in Fig. 3 are given by:

(a) $P(y|do(x), s) = \sum_z P(y|do(x), z)P(z|s)$

(b) $P(y|do(x), s) = P(y|do(x))$

(c) $P(y|do(x), s) = \sum_z P(y|do(x), z)P(z|x, s)$

Each of these formulas satisfies Theorem 1, and each describes a different procedure of pooling information from π and π^*.

For example, (c) states that to estimate the causal effect of X on Y in the target population π^*, $P(y|do(x), z, s)$, we must estimate the z-specific effect $P(y|do(x), z)$ in the source population π and average it over z, weighted by $P(z|x, s)$, i.e., the conditional probability $P(z|x)$ estimated at the target population π^*. The derivation of this formula follows by writing

$$P(y|do(x), s) = \sum_z P(y|do(x), z, s)P(z|do(x), s)$$

and noting that Rule 1 of do-calculus authorizes the removal of s from the first term (since $Y \perp\!\!\!\perp S|Z$ holds in $G_{\overline{X}}$) and Rule 2 authorizes the replacement of $do(x)$ with x in the second term (since the independence $Z \perp\!\!\!\perp X$ holds in $G_{\underline{X}}$.)

A generalization of transportability theory to multi-environment has led to a principled solution to estamability problems in "Meta Analysis." "Meta Analysis" is a data fusion problem aimed at combining results from many experimental and observational studies, each conducted on a different population and under a different set of conditions, so as to synthesize an aggregate measure of effect size that is "better," in some sense, than any one study in isolation. This fusion problem has received enormous attention in the health and social sciences, and is typically handled by "averaging out" differences (e.g., using inverse-variance weighting).

Using multiple "selection diagrams" to encode commonalities among studies, Bareinboim and Pearl (2013) "synthesized" an estimator that is guaranteed to provide unbiased estimate of the desired quantity based on information that each study share with the target environment. Remarkably, a consistent estimator may be constructed from multiple sources even in cases where it is not constructable from any one source in isolation.

Summary Result 3. *(Meta transportability) (Bareinboim and Pearl, 2013)*

- *Nonparametric transportability of experimental findings from multiple environments can be determined in polynomial time, provided suspected differences are encoded in selection diagrams.*
- *When transportability is feasible, a transport formula can be derived in polynomial time which specifies what information needs to be extracted from each environment to synthesize a consistent estimate for the target environment.*
- *The algorithm is complete i.e., when it fails, transportability is infeasible.*

6. Missing Data from Causal Inference Perspectives

Most practical methods of dealing with missing data are based on the theoretical work of Rubin (1976) and Little and Rubin (2002) who formulated conditions under which the damage of missingness would be minimized. However, the theoretical guarantees provided by this theory are rather weak, and the taxonomy of missing data problems rather coarse.

Specifically, Rubin's theory divides problems into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Performance guarantees and some testability results are available for MCAR and MAR, while the vast space of MNAR problems has remained relatively unexplored.

Viewing missingness from a causal perspective evokes the following questions:

Q1. What must the world be like for a given statistical procedure to produce satisfactory results?

Q2. Can we tell from the postulated world whether any method exists that produces consistent estimates of the parameters of interest?

Q3. Can we tell from data whether the postulated world should be rejected?

To answer these questions the user must articulate features of the problem in some formal language, and capture both the inter-relationships among the variables of interest as well as the missingness process. In particular, the model should identify those variables that are responsible for values missing in another.

The graph in Fig. 4(a) depicts a typical missingness process, where missingness in Z is explained by X and Y, which are fully observed. Taking such a graph, G, as a representation of reality, we define two properties relative to a partially observed dataset D.

Definition 4. *(Recoverability)*
A probabilistic relationship Q is said to be recoverable *in G if there exists a consistent estimate $\hat{Q}$ of Q for any dataset D generated by G. In other words, in the limit of large samples, the estimator should produce an estimate of Q as if no data were missing.*

Definition 5. *(Testability)*
A missingness model G is said to be testable *if any of its implications is refutable by data with the same sets of fully and partially observed variables.*

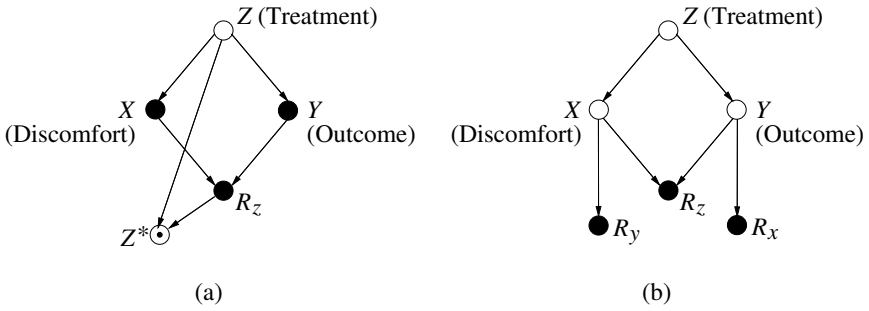

Figure 4: (a) Graph describing a MAR missingness process. X and Y are fully observed variables, Z is partially observed and Z^* is a proxy for Z. R_z is a binary variable that acts as a switch: $Z^* = Z$ when $R_z = 0$ and $Z^* = m$ when $R_z = 1$. (b) Graph representing a MNAR process. (The proxies Z^*, X^*, and Y^* are not shown.)

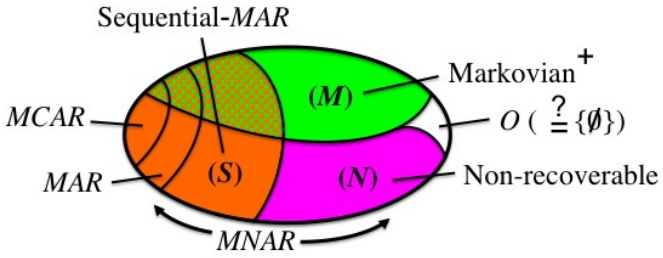

Figure 5: Recoverability of the joint distribution in MCAR, MAR, and NMAR. Joint distributions are recoverable in areas marked (S) and (M).

While some recoverability and testability results are known for MCAR and MAR, (Little, 1988; Potthoff et al., 2006) the theory of structural models permits us to extend these

results to the entire class of MNAR problems, namely, the class of problems in which at least one missingness mechanism (R_z) is triggered by variables that are themselves victims of missingness (e.g., X and Y in Fig. 4(b)). The results of this analysis is summarized in Fig. 5 which partitions the class of MNAR problems into three major regions with respect to recoverability of the joint distribution.

1. M (Markovian$^+$) - Graphs with no latent variables and no variable X that is a parent of its missingness mechanism R_x.

2. S (Sequential-MAR) - Graphs for which there exists an ordering $X_1, X_2, \ldots, X_n$ of the variables such that for every i we have: $X_i \perp\!\!\!\perp (R_{X_i}, R_{Y_i}) | Y_i$ where $Y_i \subseteq \{X_{i+1}, .., X_n\}$. Such sequences yield the estimand: $P(X) = \prod_i P(X_i | Y_i, R_{x_i} = 0, R_{y_i} = 0)$, in which every term in this product is estimable from the data.

3. N (Non-recoverable) - Graphs which are recognizable as non-recoverable. Examples are models in which X and R_x are connected by an edge or by an induced path (Verma and Pearl, 1990).

The area labeled 'O' consists of all *other* problem structures, and we conjecture this class to be empty. All problems in areas (M) and (S) are recoverable.

Note that the partition of the MNAR territory into recoverable vs. non recoverable models is query-dependent. For example, some problems permit unbiased estimation of queries such as $P(Y|X)$ and $P(Y)$ but not of $P(X, Y)$. Note further that MCAR and MAR are nested subsets of the "Sequential-MAR" class, all three permit the recoverability of the joint distribution. A version of Sequential-MAR is discussed in Gill and Robins (1997) but finding a recovering sequence in any given model is a task that requires graphical tools.

Graphical models also permit the partitioning of the MNAR territory into testable vs. nontestable models. The former consists of at least one conditional independence claim that can be tested under missingness. Here we note a peculiar property which may sound paradoxical: some testable implications of fully recoverable distributions are not testable under missingness.

Figure 4(a) demonstrates this peculiarity. Here $P(X, Y, Z)$ is recoverable since the graph is in (M) (it is also in MAR). $P(X, Y, Z)$ further advertises the conditional independence $X \perp\!\!\!\perp Y | Z$. Yet, $X \perp\!\!\!\perp Y | Z$ is not testable by any data in which the frequency of missingness (in Z) is above a certain threshold (Mohan and Pearl, 2013). Any such data can be construed as if generated by the model in Fig. 4(a), where the independence holds.

Summary Result 4. *(Recoverability from missing data) (Mohan et al., 2013)*

- *The feasibility of recovering relations from missing data can be determined in polynomial time, provided the missingness process is encoded in a causal diagram that falls in areas M, S, or N of Fig. 5.*

7. Conclusion

The unification of the structural, counterfactual and graphical approaches has produced mathematical tools that have helped resolve a variety of causal inference problems (e.g., Summary Results 1 and 2). These tools are currently being applied to new territories of statistical inference and have led to Summary Results 3 and 4.

References

Bareinboim, E. and Pearl, J. (2012a). Causal inference by surrogate experiments: z-identifiability. In de Freitas, N. and Murphy, K., editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI '12, pages 113–120, Corvalis, Oregon. AUAI Press.

Bareinboim, E. and Pearl, J. (2012b). Transportability of causal effects: Completeness results. In *Proceedings of the 26th AAAI Conference*, pages 698–704, Toronto, Ontario, Canada.

Bareinboim, E. and Pearl, J. (2013). Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 135–1434, Scottsdale, AZ.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182.

Gill, R. and Robins, J. (1997). Sequential models for coarsening and missingness. In *Proceedings of the First Seattle Symposium on Survival Analysis*, pages 295–305.

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

Halpern, J. (1998). Axiomatizing causal reasoning. In Cooper, G. and Moral, S., editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

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

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley, New York.

Mohan, K. and Pearl, J. (2013). A note on the testability of missing data models. Technical Report R-415, <http://ftp.cs.ucla.edu/pub/stat_ser/r415.pdf>, University of California Los Angeles, Computer Science Department, CA.

Mohan, K., Pearl, J., and Tian, J. (2013). Missing data as a causal inference problem. Technical Report R-410, <http://ftp.cs.ucla.edu/pub/stat_ser/r410.pdf>, University of California Los Angeles, Computer Science Department, CA. Forthcoming, NIPS-2013.

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

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

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

Pearl, J. (2000b). 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*, pages 411–420. Morgan Kaufmann, San Francisco, CA.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146. <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

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

Pearl, J. (2013). Interpretation and identification in causal mediation analysis. Technical Report R-389, <http://ftp.cs.ucla.edu/pub/stat_ser/r389.pdf>, University of California Los Angeles, Computer Science Department, CA. Forthcoming, *Psychological Methods*.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 247–254, Menlo Park, CA. AAAI Press.

Potthoff, R. F., Tudor, G. E., Pieper, K. S., and Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical methods in medical research*, 15(3):213–234.

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

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–1035.

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

Shpitser, I., VanderWeele, T., and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536. AUAI, Corvallis, OR.

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. (2002). 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.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 220–227, Cambridge, MA.

Wang, X. and Sobel, M. (2013). New perspectives on causal mediation analysis. In Morgan, S., editor, *Handbook of Causal Analysis for Social Research*, pages 215–242. Springer.