Prediction and Experimental Design with Graphical Causal Models

P. Spirtes, C. Glymour, R. Scheines, C. Meek, S. Fienberg and E. Slate

Abstract

We unify two contemporary theoretical frameworks for representing causal dependencies. Directed graphical models were introduced and developed by Kiiveri, Speed, Wermuth, Lauritzen, Pearl and others. Rubin introduced a framework for analyzing the relation between the conditional probability of Y on X and the distribution Y would have if X were forced to have a particular value. Pratt and Schlaifer have extended Rubin's analysis to offer sufficient "counterfactual" conditions for the conditional distribution of Y on Z, X= x to equal the conditional distribution of Y on Z when all units in the population are forced to have that value of X. Using two axioms for directed graphical causal models, we obtain rigorous derivations of claims given by Rubin and by Pratt and Schlaifer, and we give general characterizations in terms of causal structure-represented by directed graphs--for Pratt and Schlaifer's notions of the "observability of a law" and the "observability of a law with concomitants." Results obtained in the Rubin framework are generalized, and some relevant sampling properties of graphical causal models are obtained. [1]

Correspondence: C. Glymour, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pa. 15213. E-mail cgO9@andrew.cmu.edu

[1] Research for this paper was supported in part by contract number N00114-89-J-1964 from the Navy Personnel Research and Development Center and the Office of Naval Research.

One of the aims of an empirical study may be to predict the effects a general policy would have if put in force, or to predict relevant differences resulting from alternative policies. The interest might be in predicting the differential yield if a field is planted with one species of wheat rather than another; or the difference in number of polio cases per capita if all children are vaccinated against polio as against if none are; or the difference in recidivism rates if parolees are given $600 per month for six months as against if they are given nothing; or the reduction of lung cancer deaths in middle aged smokers if they are given help in quitting cigarette smoking; or the decline in gasoline consumption if an additional dollar tax per gallon is imposed. Such inference problems are puzzling because a policy of treatment creates a potential distribution different from the distribution sampled in observations or experiments, and alternative policies of treatment create alternative potential distributions with alternative statistics. The inference task is to move from a sample of one of these distributions, the one corresponding to passive observation or experimental manipulation, to conclusions about the distribution that would result if a policy were imposed.

A further feature makes prediction especially difficult. Empirical studies are often unable to control or randomize all of the relevant variables, with the result that the dependency among variables relevant to prediction may be confounded by unmeasured common causes. In that case, the effect of a policy that manipulates one of the variables cannot be expected to be predictable from sample statistics. There are many examples of predictions whose disappointment may be in part due to confounding. The second Surgeon General's report on smoking and health (1979, p. 43) found that mortality ratios (compared to permanent non-smokers) for those who quit smoking declined with the number of years since quitting, equaling lifelong non-smokers after 15 years. Brownlee (1965), following Fisher (1959) conjectured that such decreases might at least in part be due to self-selection of quitters caused by genetic or cultural factors. The Brownlee hypothesis can be represented by a simple picture.

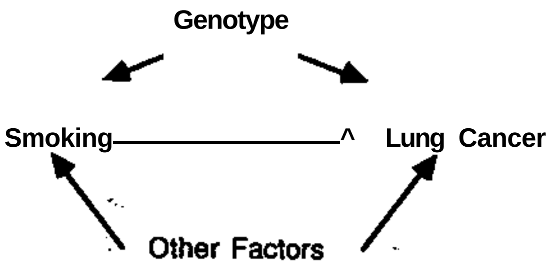

Figure 1

If accepted, the picture leads one to expect that the sample mortality ratios reported by the Surgeon General will not be good predictors of the effects of intervention against cigarette smoking. Shortly after the Surgeon General's report there appeared the results of randomized intervention studies that followed subjects for six and ten years respectively. The studies (Rose, et al., 1982; MRFIT, 1983) found that intervention that reduced cigarette smoking in middle aged males did not reduce their mortality rate.

Prediction of the effects of a policy from sample data requires knowledge of the causal processes that generated the sample. There are two questions: (i) what must be known, and (ii) when *can* it be known? Rubin (1974;1977; 1978) and following him, Holland (1986) and Pratt and Schlaifer (1984; 1988) have developed a theory that addresses the first question, and Pratt and Schlaifer (1988), have applied it to attempt to answer the second question as well. Their account has been applied to the question of randomization in experimental design (Rubin, 1978), to estimation of the difference in expected values of an outcome variable if alternative treatments were to be given to an entire population (Rubin, 1977), to determining from a causal structure when the law giving the distribution of a variable Y if a variable or set of variables X were to be forced to have a specified value is "observable," i.e., equal to the observed conditional distribution of Y on that value of X (Pratt and Schlaifer, 1988), to advice about when regression coefficients are "structural" (Pratt and Schlaifer, 1984, 1988), and to other topics.

Our aims are, first, to show how the results announced in the Rubin framework may be rigorously derived from simple axioms on graphical causal models and thereby connected with another line of statistical work on causality deriving from Kiiveri and Speed (1982), Wermuth and Lauritzen (1983), Pearl (1989) and others. We will, furthmore, generalize results in Rubin's framework and characterize in graphical terms the conditions under which their results apply.

2. Rubin's Framework and Pratt and Schlaifer's Rules

Rubin's framework has a simple and appealing intuition. In experimental or observational studies we sample from a population. Each unit in the population, whether a child or a national economy or a sample of a chemical, has a collection of properties.

Among the properties of the units in the population, some are *dispositional*--they are propensities of a system to give a response to a treatment. A glass vase, for example, may be fragile, meaning that it has a disposition to break if struck sharply. A dispositional property isn't exhibited unless the appropriate treatment is applied--fragile vases don't break unless they are struck. Similarly, in a population of children, for each reading program each child has a disposition to produce a certain post-test score (or range of test scores) if exposed to that reading program. In experimental studies when we give different treatments to different units, we are attempting to estimate dispositional properties of units (or their averages, or the differences of their averages) from data in which only some of the units have been exposed to the circumstances in which that disposition is manifested. Rubin associates with each such dispositional quantity, Q, and *each value* x of relevant treatment variable, X, a random variable, $Q_{xf,}$ whose value for each unit in the population is the value Q *would have* if that unit were to be given treatment x, or in other words if X were forced to equal x. If unit i is actually given treatment x1 and a value of Q is measured for that unit, the measured value of Q equals the value of Q_{x1f}.

Experimentation may give a set of paired values <x, y> = <x, y_{xf}>, where y_{xf} is the value of the random variable Y_{xf}. But for a unit i that is given treatment x1, we also want to know the value of Y_{x2f}, Y_{x3f}, and so on for each possible value of X, representing respectively the values for Y that unit i is disposed to exhibit if unit i were exposed to treatment x2 or x3, that is, if the X value for these units were forced to be x2 or x3 rather than x1. These unobserved values depend on the causal structure of the system. For example, the value of Y that unit i is disposed to exhibit on treatment x2 might depend on the treatments given to other units. We will suppose that there is no dependence of this kind, but we will investigate in detail other sorts of connections between causal structure and Rubin's conterfactual random variables.

A typical inference problem in Rubin's framework is to estimate the distribution of Y_{xf} for some value x of X, over all units in the population, from a sample in which only some members have received the treatment x. A number of variations arise. Rather than forcing a unique value on X, we may contemplate forcing some specified distribution of values on X, or we may contemplate forcing different specified distributions on X depending on the (unforced) values of some other variables Z; our "experiment" may be purely observational so that an observed value q of variable Q for unit i when X is observed to have value x is not necessarily the same as Q_{xf}. Answers to various problems such as these can be found in the papers cited. For example, in our paraphrasing, Pratt and Schlaifer claim the following:

When all units are systems in which Y is an effect of X and possibly of other variables, and no causes of Y other than X are measured, in order for the conditional distribution of Yon X = x to equal $Y_x f$ for all values of x of X, it is sufficient and "almost necessary" that X and each of the random variables $Y_x f$ (where xf ranges over all possible values of X) be statistically independent

Pratt and Schlaifer's principle requires us to treat *some* counterfactual variables as the same as their observed counterparts, to let the manipulated variable Xf be independent of *some* counterfactual variables, and to let *some* other counterfactual variables have their distributions be determined by manipulated variables in the same way as in the unmanipulated, sampled distribution. What principles guide these choices? What determines that some counterfactual variables are the same as their observed counterparts while others are not? Which choices are to be made in applying the framework, and why?

Pratt and Schlaifer[f]s claim may be clarified with several examples, which will also serve to illustrate some tacit assumptions in the application of the framework. Suppose X and U, which is unobserved, are the only causes of Y, and they have no causal connection of any kind with one another, a circumstance that we will represent by a diagram

X -> Y <- U

Figure 2

For simplicity we suppose the dependencies are all linear, and that for all possible values of X, Y and U, and all units, Y = X + U. Let Xf represent values of X that could possibly be *forced* on all units in the population. X is an observed variable; Xf is not. X is a random variable; Xf is not. We further simplify Pratt and Schlaiffer's set-up by giving each unit a precise value rather than a distribution of values. Consider the following table of values

Table 1

X	Y	U	X_f	$U_{Xf=1}$	$Y_{Xf=1}$
1	1	0	1	0	1
1	2	1	1	1	2
1	3	2	1	2	3
2	2	0	1	0	1
2	3	1	1	1	2
2	4	2	1	2	3

Suppose for simplicity each row (ignoring X_f, which is not a random variable) is equally probable. Here the X and Y columns give possible values of the measured variables. The U column gives possible values of the unmeasured variable U. X_f is a variable whose column indicates values of X that might be forced on a unit; we have not continued the table beyond $X_f = 1$. The $U_{Xf = 1}$ column represents the range of values of U when X is forced to have the value 1; the $Y_{Xf = 1}$ gives the range of values of Y when X is forced to have the value 1. Notice that in the table $Y_{Xf = 1}$ is uniquely determined by the value of X_f and the value of $U_{Xf = 1}$ and is independent of the value of X.

The table illustrates Pratt and Schlaifer's claim: $Y_{Xf = 1}$ is independent of X and the distribution of Y conditional on X = 1 equals the distribution of $Y_{Xf = 1}$.

We constructed the table by letting $U = U_{Xf=1}$, and $Y_{Xf=1} = 1 + U_{Xf=1}$. In other words, we obtained the table by assuming that save for the distribution of X, the causal and probabilistic structure are completely unaltered if a value of X is forced on all units. By applying the same procedure with $Y_{Xf=2} = 2 + U_{Xf=2}$, the table can be extended to obtain values when $X_f = 2$ that satisfy Pratt and Schlaifer's claim.

Consider a different example in which, according to Pratt and Schlaifer's rule, the law relating Y to X is *not* observable. In this case X causes Y and U causes Y, and there is no causal connection of any kind between X and U, as before, but in addition an unmeasured variable V is a common cause of both X and Y, a situation we will represent with the following diagram.

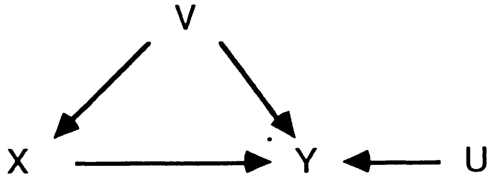

Figure 3

Consider the following distribution, with the same conventions as in Table 1:

Table 2

X	V	U	Y	Xf	$V_{Xf=1}$	$U_{Xf=1}$	$Y_{Xf = 1}$
0	0	0	0	1	0	0	1
0	0	1	1	1	0	1	2
0	0	2	2	1	0	2	3
0	0	3	3	1	0	3	4
1	1	0	2	1	1	0	2
1	1	1	3	1	1	1	3
1	1	2	4	1	1	2	4
1	1	3	5	1	1	3	5

Again, assume all rows are equally probable, ignoring the value of Xf which is not a random variable. Notice that $Y_{Xf=1}$ is now *dependent* on the value of X. And, just as Pratt and Schlaifer require, the conditional distribution of Y on X = 1 is *not* equal to the distribution of $Y_{Xf=1}$.

The table was constructed so that when X = 1 is forced, and hence Xf = 1, the distributions of $U_{Xf=1}$, and $V_{Xf=1}$ are independent of Xf. In other words, while the system of equations {Y = X + V + U; X = V} was used to obtain the values of X and Y, the assumptions $U_{Xf} = U$, $V_{Xf} = V$ and the single equation $Y_{Xf} = Xf + V_{Xf} + U_{Xf}$ were used to determine the values of $U_{Xf=1}$, $V_{Xf=1}$ and $Y_{Xf = 1}$. The forced system was treated as if it were described by the diagram:

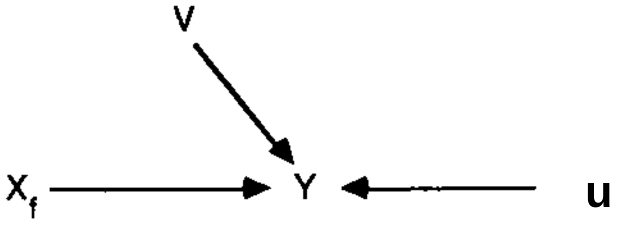

Figure 4

For another example, suppose Y = X + U, but there is also a variable V that is dependent on both Y and X, so that the system can be depicted in the following way:

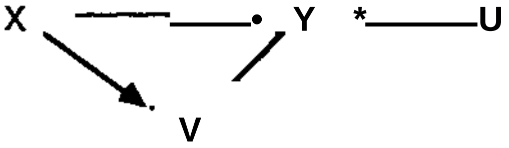

Figure 5

Here is a table of values, obtained by assuming Y = X + U and V = Y + X, and these relations are unaltered by a manipulation of X:

Table 3

X	Y	V	U	Xf	Vxf-1	Uxf-1	Yxf-1
0	0	0	0	1	2	0	1
0	1	1	1	1	3	1	2
0	2	2	2	1	4	2	3
1	1	2	0	1	2	0	1
1	2	3	1	1	3	1	2
1	3	4	2	1	4	2	3

Again assume all rows are equally probable. Note that Yxf=i is independent of X, and Yxf=1 has the same marginal distribution as Y conditional on X = 1. So Pratt and Schlaifer's principle is again satisfied, and in addition the law relating X and Y is "observable." The table was constructed by supposing the manipulated system satisfies the very same system of equations as the unmanipulated system, and in effect that the diagram of dependencies in figure 4 is unaltered by forcing values on X.

Consider finally an example due to Rubin. In an experiment in which treatments T are assigned on the basis of a randomly sampled value of some variable X which shares one or more unmeasured common causes, V, with Y, we wish to predict the average difference *x* in Y values if all units in the population were given treatment T = 1 as against if all units were given treatment T = 2. The situation in the experiment can again be represented by a diagram:

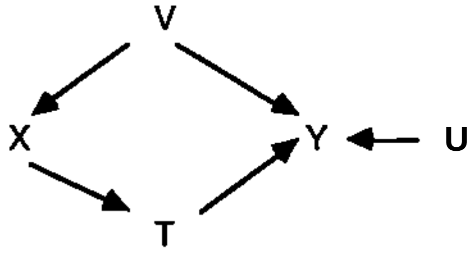

Figure 6

The pre-test (X), reading program assignment (T), post-test (Y) case is a model. Provided the experimental sample is sufficiently representative, Rubin says that an unbiased estimate of *x* can be obtained as follows: Let k range over values of X, let $\overline{Y1k}$ be the average value of Y conditional on X = k and T = 1, and analogously for $\overline{Y2k}$. Let n1k be the number of units in the sample with T = 1 and X = k, and analogously for n2k. The numbers n1 and n2 represent the total number of units in the sample with T =1 and T =2 respectively. Then estimate the expected value of Y if treatment 1 is forced on all units by

$$\sum_{k=1}^{K} \frac{n1k + n2k}{n1 + n2} \overline{Y1k}$$

and estimate *x* by:

$$\sum_{k=1}^{K} \frac{n1k + n2k}{n1 + n2} [\overline{Y1k} - \overline{Y2k}]$$

The basis for this choice may not be apparent. In the distribution of experimental units, the expected value of Y *conditional on T = 1* is

$$\Sigma_k \; P(X \mid T = 1) \; P(Y \mid T = 1, X = k).$$

But in calculating the expected value of Y if *treatment 1 is forced on all units,* Rubin substitutes the formula

$$\Sigma_k \; P(X\ ) \; P(Y \mid T = 1, X = k)$$

Now, P(X | T) is equal to P (X) if and only if X and T are independent. In other words, in calculating the distribution of Y when T is forced to have value 1 (or when forced to have value 2), Rubin treats X and T as independent. In effect, he assumes the manipulated systems would have the structure:

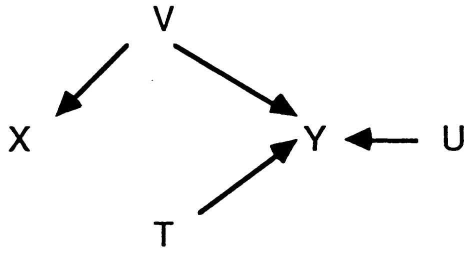

Figure 7

We trust these examples help to clarify the idea behind Pratt and Shlaifer's principle, but they don't themselves answer the chief questions: Pratt and Schlaifer's principle is satisfied when (1) we treat *some* counterfactual variables as the same as their observed counterparts, (2) let the manipulated variable Xf be independent of *some* counterfactual variables, and (3) let *some* other counterfactual variables have their distributions determined by Xf, U_{Xf} and V_{Xf} in the same way that the distribution of Y is determined by X, U and V. What principles guide these choices? If, to modify Rubin's example, treatment were assigned not on the basis of pre-test X but on the basis of some variable such as family income which presumably is a cause of X, could we predict the conditional distribution of Y on X values if all children were given a particular reading program? Presumably there is some principle that determines which variables are changed in their marginal distributions upon a manipulation, which are not changed, and which formerly dependent variables are made independent. We will give such a principle in the context of graphical causal models.

3. Graphical Causal Models

Under appropriate conventions, the diagrams we have drawn are not only depictions of hypothetical causal dependencies, they are also perfectly definite statistical hypotheses. We will assume that causal relations among variables are represented by a directed acyclic graph. The variables of interest are vertices of the graph and a directed edge X -> Y indicates that X is a cause of Y whose influence is not entirely mediated by other variables in the graph. Following Kiiveri and Speed (1982) and Pearl (1988) we will assume that if units satisfying such causal relations are to be sampled randomly the distribution thus determined satisfies the Markov condition, or more strongly, the Faithfulness condition.

Markov Condition: A graph G and a probability distribution P on the vertices V of G satisfy the Markov condition if and only if for every V in V, and every set X of vertices in V such that no member of X is a descendant of V , V and X are independent conditional on the parents of V.

Faithfulness Condition: P is faithful to G provided for all disjoint sets of variables X, Y, and Z, X and Z are independent conditional on Y in P if and only if X and Z are independent conditional on Y in every distribution satisfying the Markov condition for G.

The Faithfulness condition says that the Markov condition applied to G completely axiomatizes the conditional independence relations in P. For linear causal systems with multinormally distributed variables it can be shown that the set of linear coefficients and exogenous variances for which the Markov condition is satisfied by a graph but faithfulness is not satisfied has zero Lebesque measure. As we wifl see, Pratt and Schlaifer assume a condition that is at least as strong as Faithfulness.

For any distribution satisfying the Markov condition on a graph G with vertex set V, the joint density can be factored

$$\prod_{i=1}^{n} P(V_i \mid \pi_i)$$

where π_i denotes the (possibly empty) set of parents of vertex V_i (Kiiveri and Speed, 1982). Directed independence graphs represent hypotheses about marginal independence and marginal conditional independence relations as well as non-marginal relations. We will use "conditional independence" to include both marginal and non-marginal relations. For multinomial distributions the maximum likelihood estimate of the distribution subject to the conditional independence constraints represented by the causal graph can be obtained by substituting the marginal sample frequencies into formula (I).

In keeping with the discussion of section 1, we are interested in making predictions about distributions that are produced by forcing a distribution on a variable(s) in a way that breaks causal links *into* that variable, and only such links, in the causal structure actually sampled. We make one assumption about distributions produced by such forcing:

Manipulation Condition: If the original distribution is

$$\prod_{i=1}^{n} P(V_i|\pi_i)$$

then a manipulation of X to force P_X on X--in such a way that changes in distributions of other variables are due entirely to changes in the distribution of X--gives the same distribution save that $P(X|\pi_X)$ is replaced by P_X.

In view of the factorization principle for probability distributions satisfying the Markov condition, the Manipulation Condition implies that in a system originally represented by graph G, the directed graph G' representing the system upon forcing a distribution P(X) is obtained by deleting from G all of the edges directed into X. A generalization of the Manipulation Condition says that if we force a distribution on X depending on the unforced values of a variable Z, the original conditional distribution of X on its parents should be replaced by the forced conditional distribution of X on Z, and the original graph replaced by a graph that is otherwise the same save that there is an edge from Z into X and there are no other edges directed into X.

In each of the examples illustrating Pratt and Schlaifer's condition the tables of values accord with the Manipulation Condition--indeed that is how we calculated them. Likewise, Rubin's analysis of treatment assignment determined by a covariate gives exactly the result that would be obtained by applying the Manipulation Condition. That seems to us good reason to think that the structure the Rubin framework is after is caught by the Markov, Faithfulness and Manipulation Conditions. A "model-free" theory of causal inference and experimental design follows from these axioms, and in practice the entire story is contained in the diagrams. If the causal structure of the experimental system is known, determining observability is as simple as drawing a picture and then erasing one or more lines.

While we will show how to obtain equivalents in graphical terms of Rubin's and Pratt and Schlaifer's results, it should be pointed out that the framework is more general. The results we have described in the Rubin framework give rules for predicting particular features of a joint distribution provided the causal structure is known and satisfies various conditions. The Markov and Manipulation conditions specify, for any causal structure given by an acyclic directed graph, any appropriate joint distribution, and any distribution or conditional distribution to be imposed on any subset of the variables, the resulting joint distribution of all of the variables.

4. Sampling and Conditional Probabilities

If the conditional probability relations for variables in units with the same structure are known, then by the Manipulation Condition we can derive the probability distribution that will be obtained for any variable Y upon forcing a distribution on a set of variables X in a fashion that satisfies the antecedent of the Condition. Hence an estimate of the conditional probability relations leads to a prediction. In both observational and experimental studies we sample from a population; the difference is that in experimental studies we force a distribution on some of the variables. Rubin (1978) emphasizes, correctly we believe, the importance of sample selection and assignment mechanisms. However we sample, to apply the Manipulation Condition to predict the effects on variable Y of manipulations or policies directed to variables X, we need to know if the conditional probabilities can be consistently estimated from the sample. Suppose there is a population whose units are each described by a directed graph G of causal structure; let the values of the variables be distributed as P faithfully to G. Under what methods of sampling will the conditional probabilities for sampled variables be as in P?

We will assume that a sample is always obtained by specifying a property S that, like other variables of concern, has a distribution of values in the population. In the simplest case S can be viewed as a binary variable with the value 1 indicating that a unit has the sample property. In principle one might wish to select a sample by using a variable with several values (e.g., age group) and drawing fixed numbers or proportions from each group. So our general questions concern when conditioning on any value of S leaves unaltered the conditional probabilities or conditional independence relations for variables in the vertex set of G. We will not consider questions about the sampling distributions obtained by imposing various constraints on the distribution of values of S in a sample. Our treatment assumes that S is not logically connected with any of the variables in G, but trivial variations of the theorems apply when S is identical with one of those variable.

The graph G can be expanded to a graph G(S) that includes S and whatever causal relations S and the other variables realize. We assume a distribution P(S) faithful to G(S) whose marginal distribution summing over S values will of course be P. We suppose that the sampling distribution is determined by the conditional distribution P(_|S). Our questions are then, more precisely, when this conditional distribution has the same conditional probabilities and conditional independence relations as P. We require, moreover, that the answer be given in terms of the properties of the graph G(S). To give the full answer we require some simple graph theoretic definitions and a lemma.

Any sequence of vertices joined by edges of any orientation, e.g. A->B<-C<-D->E, is an *undirected* path, and any vertex in which two edges meet in such a path is a *collider* on the path, e.g., B in the path illustrated. An undirected path is said to be *into* a terminus, e.g. E, if it contains an edge into E, and *out of* a terminus, e.g., A, if it contains an edge out of A. A *parent* of vertex V in graph G is any vertex U such that U->V in G; a *descendant* of U is any vertex V (including U) such that there is a directed path from U to V.

In directed graph G, variables X, Y are *d-connected* relative to a set **Z** of variables not containing X or Y provided there exists a sequence p of edges (an undirected path) connecting X, Y such that no non-collider on p is in **Z** and every collider on **Z** has a descendant in **Z**. (Pearl, 1988; Lauritzen, et al. 1990). Any undirected path with the properties of p is said to be a *d-connecting path* for X, Y with respect to **Z**. X and Y are *d-separated* by **Z** if and only if X and Y are not d-connected given **Z**. For any three disjoint sets of variables **X**, **Y**, and **Z**, **X** and **Y** are d-separated by **Z** if and only if every member of **X** is d-separated from every member of **Y** by **Z**.

Assume for the moment that all variables are discrete.

Lemma: For any directed acyclic graph G and probability distribution P on the vertices of G, P is faithful to G if and only if for all vertices X, Y and every set **Z** of vertices, **Z** d-separates X, Y if and only if X, Y are conditionally independent on **Z**.

That d-separability characterizes the Markov condition was proved by Verma. See Pearl (1988).

Theorem 1 Let X, S, Y be distinct discrete variables in G, P a distribution faithful to G. Then $Y \perp\!\!\!\perp S \mid X$ if and only if {X} d-separates Y and S in G(S).

(We use Dawid's notation, so that $Y \perp\!\!\!\perp S \mid X$ signifies that Y, S are independent conditional on X). The theorem follows immediately from the Lemma since for discrete variables P(Y | X) = P(Y | X, S) if and only if Y, S are independent conditional on X. A parallel lemma can be proved for linear systems with partial correlations substituted for conditional independence, and a corresponding theorem then follows (Spirtes, 1991b).

Theorem 1 generalizes to any number of conditioning variables X1,...,Xn. The theorem is essentially the observation that for any covariate Z, P(Y | XZ) = P(Y | XZS) if and only if in P the variables Y and S are independent conditional on {XZ}. It entails, for example, that if we wish to estimate the conditional probability of Y on X from a sample of units with an S property (say, S = 1) and a Z property, we should try to ensure that there is

(i) no direct connection between Y and S,
(ii) no trek between Y and S that does not contain X or Z, and
(iii) no pair of treks between Y and Z and between S and Z that are both into Z, and similarly for Y and X.

Our sampling property should not be the direct or indirect cause or effect of Y save through a mechanism that is blocked by holding X and Z constant, and neither X nor Z should not be the effect, direct or indirect of both Y and the sampling property. (The latter clause in effect guarantees that Simpson's paradox is avoided in a faithful distribution).

These examples are of course not exhaustive. Theorem 1 entails a partial justification of the conventional wisdom: "prospective" sampling is more reliable than "retrospective" sampling if by the former is meant a procedure that selects by a property associated with Y, the effect, only through X, the cause, and by the latter is meant a procedure that selects by a property associated with X only through Y.

5. **Prediction and Experimental Design**

The Markov and Manipulation Conditions provide the basis for prediction from samples obtained by experiment, quasi-experiment or observation provided appropriate facts about the causal structure of the sampled population is known. We will consider several issues of design taken up by Rubin or by Pratt and Schlaifer.

5.1 Randomization

Suppose as before that the goal is to estimate the conditional probability P(Y | X) in distribution P. In drawing a random sample of units from P we in effect sample according to a property S that is entirely disconnected, graphically, from the variables of interest in the system. If we succeed in doing that then we ensure that S has no causal connections that can bias the estimate of the conditional probability. Of course because of sample variation an actual sample so obtained might give a poor estimate of the conditional probability in the population. Substantive knowledge about the causal relations of the sampling property could in principle substitute for randomization.

5.2 Observability

Pratt and Schlaifer are concerned with the following question: If X and Y are measured, and no other relevant variables are measured, when is the conditional distribution of Y on any value x of X equal to the distribution Y would have if X were forced to have value x? If the equality holds they say the law relating the distributions of Y and X is "observable." One of their "sufficient and almost necessary" conditions for observability is that X and Y_xf be independently distributed for any value xf of Xf. The condition can only be applied if we know the joint distribution of X and Y_xf. The Manipulation Condition gives the following result:

Theorem 2: Assuming the Faithfulness and Manipulation Conditions, and assuming that all conditional probabilities are positive, the law giving the dependency of Y on X will be "observable" if and (almost) only if in the sampled systems no undirected path *into* X d-connects X, Y with respect to the empty set of vertices. Equivalently, if and (almost) only if (1) Y is not a cause of X, and (2) there is no common cause of X and Y.

The sufficiency claimed in Theorem 2 is a special case of Theorem 3 below and requires only the Markov and Manipulation Conditions. "Necessity" is used in the vague sense of Pratt and Schlaifer's claim: unless the distribution satisfies special constraints, the condition is necessary assuming Faithfulness.

In comparing the conditional distribution of Y on X in our experimental or observational sample with the distribution of Yxf we are comparing factors in distributions for two graphs. The original graph, G, represents the causal structure of the observed or experimental system; the other graph, G\ is the original graph, *minus all of the edges directed into X.* Yxf will be distributed in the same way as Y conditional on X if and only if when we use the factorization formula (I) on the two graphs, we find in each case the same formula for the conditional distribution of Y on X. Theorem 2 gives the sufficient graphical condition for the sameness of the conditional distribution of Y on X in G and in G^f.

Theorem 2 is illustrated in the three tables of section 2 and the accompanying causal graphs. For example if G is given by figure 2, then G^f is given by the same figure, and the conditional distribution of Y on X is trivially the same in both cases. If G is given by the graph in figure 3 , then G^1 is given by figure 4 and the conditional distribution of Y on X for G is $X_{UV}P(Y \mid X,U,V)P(X \mid V)P(U)P(V)$ while for G^1 it is $X_{UV}P(Y$ I X,U,V)P(X)P(U)P(V) where probabilities of the same arguments have the same values for G and G^1. Save for odd cases, the two expressions are equal only if $P(X) = P(X \mid V)$, which will not be the case if P is a distribution faithful to G. The Faithfulness Condition captures part of what is meant by Pratt and Schlaifer^{f}s claim that their rule is "almost necessary."

5.2 Covariates and Predicting Conditional Distributions in Manipulated Populations

Pratt and Schlaifer consider the case in which, besides X and Y, some further variables **Z** are measured. Their discussion is a generalization of Rubin's (1977) discussion of experiments in which treatment assignments depend on some variable that is a cause of the outcome variable or shares a common cause with the outcome variable. (Rubin's X is Pratt and Schlaifer's Z; Rubin's T is Pratt and Schlaifer's X; we will keep with Pratt and Schlaifer's notation.) We know from Theorem 2 that if in the sampled or experimental systems Z is a cause of both X and Y, or a cause of X and shares a common cause with Y, then the distribution of Y if a value x is forced on X will not be the same as the conditional probability of Y on X. The dependency of Y on X will not be "observable." Rubin's observation is in effect that in this situation we can nonetheless use the conditional probability of Y given Z = z and X = x to estimate the distribution of Y given Z = z when X is forced to have the value x. Pratt and Schlaifer say in this case that the law relating Y to X is "observable with concomitant Z." With an estimate in hand for each value z of Z of the distribution of Y conditional on Z = z when X is forced to have the value x, by summing over the values of the concomitant Z, Rubin estimates the average value Y when X is forced to have the value x. Pratt and Schlaifer also claim sufficient and "almost necessary" conditions for observability with concomitants, namely that for any value x of X the distribution of X be independent of the conditional distribution of Y_X on the value of z of Z_{Xf} when X is forced to have the value x. We will give a general sufficient condition in graphical terms, a condition which follows necessarily from the Markov and Manipulation Conditions.

Theorem 3: Assuming the Markov and Manipulation Conditions and positivity for all conditional probabilities, for all values of X and of **Z**, the distribution of Y conditional on X = x and **Z** equals the distribution of Y conditional on **Z** when X is forced to have the value x if no path d-connecting X, Y relative to **Z** is *into* X.

The proof is given in the Appendix.

Pratt and Schlaifer say their condition is "almost necessary." What they mean, we take it, is that there are cases in which their condition fails to hold but they arise only if a special constraint is satisfied by the conditional probabilities. Parallel remarks apply to the graphical condition given in Theorem 3.

6. Causal Inference

Some disputes are less about how much one variable affects another and more about whether one variable affects another at all. For example in the 1950s R. A. Fisher and many epidemiologists were not so much concerned with estimating how much of the incidence of lung cancer was due to smoking, or how much the rate of lung cancer would decline if people stopped smoking; they were principally concerned with whether smoking has a causal role in producing lung cancer. (Cook, 1979). Consider questions of the following kind: Does X influence Y at all? Does X influence Y by any mechanism that cannot be blocked by controlling Z? Answers to such questions are to be obtained from a sample of units that may or may not have been subjected to some experimental treatment. As before, the first issue is to characterize sampling procedures that do not bias answers to such questions.

In Theorem 4 let Z be any set of variables in G not including X and Y.

Theorem 4: Exactly one of < XJLY I Z; XJLY | Z U {S}> is false if and only if the corresponding member and only that member of <Z d-separates X, Y; Z U {S} d-separates X, Y> is false.

Theorem 4 is an obvious application of the previous Lemma. Suppose in the ambient distribution P that X and Y are independent conditional on Z. When will sample property S make it appear that X and Y are instead dependent conditional on Z? The answer is exactly when X, Y are dependent conditional on Z U {S} in P(S). This circumstance-conditional independence in P and conditional dependence in P(S)-can occur for faithful distributions when and only when there exists an undirected path q from X to Y with special properties. Then for conditional independence in P and conditional independence in P(S) there must exist an undirected path q in the causal graph such that (i) no non-collider on q is in Z U {S}; (ii) every collider on q has a descendant in Z U {S}; and (iii) some collider on q does not have a descendant in Z.

The converse error involves conditional dependence in P and conditional independence in P(S). That can happen in a faithful distribution when and only when there exists an undirected path U from X to Y such that (i) every collider on U has a descendant in Z; and (ii) no non-collider in U is in Z, and S is a non-collider on every such path. Again, asymptotically both of these errors can be avoided by sampling randomly, that is by a property S that is causally unconnected with the variables of interest.

7. Conclusion

The Rubin framework for comparing distributional properties between manipulated and unmanipulated (or differently manipulated) populations is a first step towards a principled understanding of prediction and experimental design, and we hope the derivation of the claims of that framework from axioms on directed graphical models represents a further step towards the same goal. It is important to recognize, however, how small these steps are, and how far they leave us from the goal.

The property Pratt and Schlaiffer call "observability" is better termed *conditional probability invariance under manipulation*, since their principles are neither necessary nor sufficient for predicting features of a distribution of manipulated systems from distributional properties of a population of unmanipulated systems. The insufficiency is due to the fact that Pratt and Shlaiffer's and Rubin's conditions presume counterfactual--that is, in most cases, causal--knowledge about the systems under study, but provide no information as to when the marginal distribution of measured variables and background knowledge determine enough about causal structure to conclude that particular conditional probabilities are the same in the manipulated and unmanipulated distributions. Assuming either the Markov or the Faithfulness conditions, there are connections between causal structure and marginal distributional structure; the implications of these connections for conditional probability invariance under manipulation are described in Spirtes (1991b).

That conditional probability invariance is unnecessary for the prediction of features of manipulated distributions can be illustrated in a simple linear case. Suppose the causal structure is

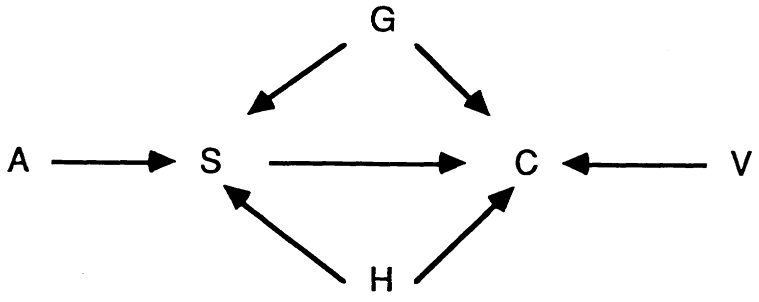

Figure 8

and all dependencies are linear. Suppose only A, S and C are observed. For concreteness, we can interpret A as cigarette advertising exposure, S as cigarette smoking, C as age at death, although it is not very likely these variables are linearly related. Let b be the linear coefficient of S in the equation for C. For a standardized model, b equals the partial correlation of C, S controlling for all unmeasured causes. If S is manipulated according to the Manipulation Condition, the graph is the same save that the G -> S and H -> S edges are omitted, the linear equations are correspondingly altered, and S becomes independent of A, G and H. The conditional density of C given S is not the same in the manipulated and unmanipulated distributions, and the "law" relating S and C is not "observable," in Pratt and Schlaiffer's terminology.

Now suppose one knew the following about the causal structure: A has no effect on C save possibly through S. Then b is determined by the ratio of the correlation of A and C to the correlation of S and C. But b determines the difference in mean values of C for two manipulations that assign different values to S. Asymptotically, given the assumption of linearity and the substantive knowledge that A has no effect on C save through S, the differental effects of alternative smoking behaviors on mortality could be predicted without any knowledge of the unobserved common causes.

Finally, it should be noted that while the Manipulation Condition captures the intuitions of the Rubin framework in graphical terms, in many cases it is not a correct representation of the relation between a distribution of unmanipulated units and the distribution that would result from a manipulation. In the Rose and MRFIT studies of smoking and mortality, for example, the Manipulation Condition arguably did not apply because the counseling and support treatment given to one group merely added another cause of smoking behavior in the members of that treatment group. Other causes of smoking behavior in the treatment and non-treatment groups were not disengaged, as they might have been the case if, for example, members of one group had been forced not to smoke and members of the other had been forced to continue smoking.

References

Asmussen, S. and Edwards, D. 1983. Collapsibility and response variables in contingency tables. Biometrika, 70, 567-578. v-

Bishop, Y, Fienberg, S, and Holland, P. 1975. Discrete Multivariate Analysis: Theory and Practice. M.I.T. Press. Cambridge.

Brownlee, K. 1964. A review of "Smoking and Health" Journal of the American Statistical Association, 60, 722-739.

Cook, R.D. 1980. Smoking and lung cancer, in S. Fienberg and D. Hinkley, (eds.), R.A. Fisher: An Appreciation. Springer-Verlag, Berlin.

Fisher, R. 1959. Smoking. The Cancer Controversy. Some Attempts to Assess the Evidence. Edinburgh, Oliver and Boyd.

Greenland, S. 1989. Modeling and variable selection in epidemiologic analysis. American Journal of Public Health 79. 340-349.

Holland, P. 1986. Statistics and causal inference. Journal of the American Statistical Association 81. 945-960.

Kiiveri, H., Speed, T.1982. Structural analysis of multivariate data: a review. In Leinhardt, S. (ed.) Sociological Methodology. Jossy Bass: San Francisco.

MRFIT Research Group. 1982. Multiple risk factor intervention trial: risk factor changes and mortality results. Journal of the American Medical Association 248, 1465-1477.

Pearl. J. 1988. Probabilistic Reasoning in Intelligent Systems. Morgan and Kaufman: San Mateo.

Pratt, J. and Schlaifer, R. 1984. On the nature and discovery of structure. Journal of the American Statistical Association 79. 9-21.

Pratt, J. and Schlaifer, R. 1988. On the interpretation and observation of laws. Journal of Econometrics, 39. 23-52.

Rose, G., Hamilton, P., Colwell, L. and Shipley, M. 1982. A randomized controlled trial of anti-smoking advice: 10-year results. Journal of Epidemiology and Community Health, 36, 102-108.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66. 688-701.

Rubin, D. 1977. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 2. 1-26.

Rubin, D. 1978. Bayesian inference for causal effects. Annals of Statistics 6. 34-58.

Simpson, E. 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, B, 13, 238-241.

Spirtes, P., and Glymour, C. 1988. Latent variables, causal models and overidentifying constraints. Journal of Econometrics, 39. 175-198.

Spirtes, P., Glymour, C., and Scheines, R. 1991a. An algorithm for fast recovery of sparse causal graphs." Social Science Computer Review, 9, 62-72.

Spirtes, P., Glymour, C., and Scheines, R. 1991b. Causality, Prediction and Search. Springer-Verlag, forthcoming.

Surgeon General of the United States. 1979. Smoking and Health. U. S. Government Printing Office.

Verma, T. and Pearl, J. 1990. Equivalence and synthesis of causal models. Proceedings of the Conference on Uncertainty in Artificial Intelligence, July 1990.

Wermuth, N. and Lauritzen, S. 1983. Graphical and recursive models for contingency tables." Biometrika, Vol. 72, pp. 537-552.

Whittaker, J. 1990. Graphical Methods in Applied Multivariate Statistics. Wiley, New York.

Appendix

Let P be the distribution over a set of variables containing X in a population, and P' be the corresponding distribution when X is manipulated to have the value x. Y is **observable conditional on Z, when X is manipulated** iff P(Y|Z) = P'(Y|Z). We will show that if P is faithful to the graph G of the causal process that generated P then Y is observable conditional on XZ when X is manipulated if in G there are no paths that d-connect X to Y given Z that are into X, and Y is observable conditional on Z when X is manipulated if in G, X either has no descendants in YZ or X is not d-connected to Y given Z.

In a directed acyclic graph G containing Y and Z, if Y is not in Z, then V is in **IV(Y,Z)** (informative variables for Y given Z) if and only if V is not in Z, V is d-connected to Y given Z, and V has a descendant in YZ. W is in **IP(Y,Z)** (W has a parent who is an informative variable for Y given Z) if and only if W is a member of Z, and W has a parent in IV(Y,Z).

Lemma 1: If G is a directed acyclic graph over V, P is a distribution faithful to G, and V **=V\ND(YZ)** then

$$P(Y\backslash Z) = \frac{\underset{IV(KZ)}{£} \quad \underset{w_{*}\ IV(y>Z)\ \cup\ \mathbf{IP(Y,Z)} \cup \{Y\}}{\Pi} \quad P(W|\pi_W)}{\underset{\mathbf{IV(V,Z)\ u\ \{V\}}\ \mathit{We}\ \mathbf{IV(V,Z)\ u\ IP(V,Z)\ u}\ \mathit{\{Yj}}{\mathbf{n}} \quad P(W|\pi_W)}$$

Lemma 2: If G is a directed acyclic graph containing X, Y, and Z, X and Y are not in Z, X * Y, and G^1 the corresponding graph when X is manipulated, and V * X, and no path that d-connects X to Y given Z is into X, then V is in **IV(Y,XZ)** in G if and only if V is in IV(Y,XZ) in G^f.

Proof. It is trivial that if V is in **IV(Y,XZ)** in G^f then V is in IV(Y,XZ) in G_f because G^1 is a subgraph of G.

Suppose then that V * X, no path that d-connects X to Y given Z is into X, and V is in IV(Y,XZ) in G. By definition, V is not in Z. Then in G there is a path that d-connects V to Y given **XZ,** and V has a descendant in XYZ. We will show that in G^1 there is a path that d-connects,V to Y given XZ.

Suppose first that there is a path U in G that d-connects Y to V given XZ that does not contain an edge into X. It follows that there is a corresponding path U^1 in G^1. Every non-collider on U^1 is not in Z. If every collider on U^1 has a descendant in XZ then U^1 d-connects V and Y given Z in G\ Otherwise let R be the collider on U^1 closest to Y that does not have a descendant in XZ in G'; in G every directed path D from R to a member of XZ contains X. Let Q be the point of intersection of D and the subpath of U from R to Y that is closest to Y on U. The concatenation of the subpath of D from Q to X and the subpath of U from Q to Y is into X and d-connects X and Y given Z in G, contrary to our assumption.

Suppose then that every path U in G that d-connects V to Y given XZ is into X. It follows that X is a collider along U. If there is no collider along the subpath of U from X to Y such that every directed path from the collider to a member of Z contains X, then in G the subpath of U from X to Y d-connects X to Y given Z, and is into X, contrary to our assumption. Otherwise this reduces to the previous case.

We will now prove that if V is in IV(Y,XZ) in G then V has a descendant in XYZ in G^1. Suppose that V is in IV(Y,XZ). V has a descendant in XYZ in G. Suppose that V does not have a descendant in XYZ in G^1. It follows that in G every directed path from V to a member of XYZ contains X. Because in G there is directed path from V to some member of XYZ, in G there is a directed path from V to X that contains no member of Z. Let U1 be a directed path from V to X that contains no member of Z in G. We have already proved that V and Y are d-connected given Z in G\ so there is some path U2 that d-connects V and Y in G that does not contain X, such that every collider on U2 is the source of a directed path to a member of Z that does not contain X. Let Q be the vertex on U2 closest to Y such that Q is on both U1 and U2. The concatenation of the subpath of U1 from Q to X and the subpath of U2 from Q to Y is a path that d-connects X and Y given Z in G_f and is into X, contrary to our assumption. Q.E.D.

Theorem 3: If P is a distribution faithful to the directed acylic graph G of the causal process that generated P, G contains X, Y, and Z, X and Y are not in Z, X * Y, , G^f is the graph resulting from manipulating X, and P^1 the distribution resulting from the manipulation of X, then Y is observable conditional on XZ when X is manipulated if in G there are no paths that d-connect X to Y conditional on Z that are into X.
Proof. Suppose G is a directed acylic graph, G^f is the graph resulting from manipulating G, and in G there are no paths that d-connect X and Y with respect to Z that are into X.

By lemma 2, **IV**(Y,**XZ**) is the same in G and G'.

Because the parent relationship is the same in G and G' except for the parents of X, it follows that **IP**(Y,X**Z**) is the same in G and G' with the possible exception of X. X is not in **IP**(Y,X**Z**) in G' because X has no parents in G'. Suppose that X is in **IP**(Y,X**Z**) in G. It follows that X has a parent V not in X**Z** that is in **IV**(Y,X**Z**) in G. Hence V is d-connected to Y given X**Z** in G by some path U. If U contains X then X is a collider on U because otherwise U does not d-connect Y and V given X**Z**. It follows then that X is d-connected to Y given **Z** by the subpath of U from X to Y that is into X, contrary to our assumption. If U does not contain X then the concatenation of U with the edge from V to X is a path from X to U that d-connects X and Y given **Z** that is into X, contrary to our assumption. Hence X is not in **IP**(Y,X**Z**) in G, and **IP**(Y,X**Z**) is the same in G and G'.

By hypothesis, $P(W|\pi_W) = P(W|\pi_W)$ for all W in **V**, except for W = X. By lemma 1, P(Y|X**Z**) = P'(Y|X**Z**). Q.E.D.

Lemma 3: If G is a directed acyclic graph containing X, Y, and **Z**, X and Y are not in **Z**, X ≠ Y, and G' is the corresponding graph when X is manipulated, X is not in **IV**(Y,**Z**), and V ≠ X then V is in **IV**(Y,**Z**) in G if and only if V is in **IV**(Y,**Z**) in G'.

Proof. It is trivial that if V is in **IV**(Y,**Z**) in G' then V is in **IV**(Y,**Z**) in G, because G' is a subgraph of G.

X is not in **IV**(Y,**Z**) in either G or G'. Suppose then that V ≠ X, and V is in **IV**(Y,X**Z**) in G. First we will show that V and Y are d-connected given **Z** in G'; then we will show that if V and Y are d-connected given **Z** in G' then V has a descendant in Y**Z** in G'. It will follow that V is in **IV**(Y,**Z**) in G'.

First we will show that V is d-connected to Y given **Z** in G'. Suppose, contrary to the hypothesis that V is not d-connected to Y given **Z** in G'. In G, either there is a path U(V,Y) d-connecting V and Y given **Z** that contains an edge into X, or there is some path U(V,Y) d-connecting V and Y given **Z** contains a collider W for which every directed path to a member of **Z** contains X.

Suppose first that some path U(V,Y) d-connecting V and Y given **Z** contains a collider W for which every directed path to a member of **Z** contains X. Let U(W,X) be the directed path from W to X, and U(W,Y) be the subpath of U(V,Y) from W to Y. If U(W,X) does not intersect U(W,Y) except at W then the concatenation of U(W,X) and U(W,Y) d-connects X and Y given Z. If U(W,X) does intersect U(W,Y) at some vertex not equal to W, let Q be the vertex on U(W,Y) closest to Y where they intersect. Let U(Q,Y) be the subpath of U(W,Y) from Q to Y, and U(Q,X) the subpath of U(W,X) from Q to X. Q is not in Z because no member of Z is on U(W,X). The concatenation of U(Q,X) and U(Q,Y) is an undirected path that d-connects X and Y given Z in G^f.

Suppose next that there is a path U(V,Y) d-connecting V and Y given Z that contains an edge into X. Let U(X,Y) be the subpath of U(V,Y) from X to Y. If there is a collider W on U(X,Y) such that every directed path from W to a member of Z contains X, this reduces to the previous case. Hence U(X,Y) d-connects X and Y given Z.

Next we will show that if V is d-connected to Y given Z in G\ then V has a descendant in YZ. Because V is d-connected to Y given Z in G^f by some path U(V,Y), either U is into V, or V has Y or a collider on U as a descendant. Every collider on U has a descendant in Z, so if U is not into V, then V has a descendant in YZ. Suppose then that U(V,Y) is into V. In G, V has a descendant in YZ. Hence V has a descendant in YZ in G^1 unless every directed path from V to a member of YZ contains X. In G, there exists a path from V to a member of YZ; if in addition every directed path from V to a member of YZ contains X it follows that X has a descendant in YZ in G. Let U(V,X) be a directed path from V to X. If U(V,X) does not intersect U(V,Y) then the concatenation of U(V,X) and U(V,Y) d-connects X and Y given Z. If U(V,X) does intersect U(V,Y) let Q be the vertex on U(V,Y) closest to Y that is a point of intersection. Let U(Q,X) be the subpath of U(V,X) from Q to X, and U(Q,Y) be the subpath of U(V,Y) from Q to Y. Then the concatenation of U(Q,X) and U(Q,Y) d-connects X and Y given Z. Hence X is in IV(Y,Z), contrary to our assumption. Q.E.D.

Theorem 4: If P is a distribution faithful to the directed acylic graph G of the causal process that generated P, G contains X, Y, and Z, X and Y are not in Z, X * Y, , G^1 is the graph resulting from manipulating X, and P^f the distribution resulting from the manipulation of X_f then Y is observable conditional on Z when X is manipulated if X is not in **IV(Y,Z).**

Proof. Suppose G is a directed acylic graph, G^1 is the graph resulting from manipulating G, and in G, and X is not in **IV(Y,Z).**

By lemma 3_f **IV(Y,Z)** is the same in G and G'.

Because the parent relationship is the same in G and G^1 except for the parents of X and X is not in Z, it follows that IP(Y,Z) is the same in G and G^f.

By hypothesis, P(W|TCW) = P(W|rcw) for all W in V, except for W = X. By lemma 1, P(Y|XZ) = P'(Y|XZ). Q.E.D.