

THEORY AND METHODS

An overview of relations among causal modelling methods

Sander Greenland^a and Babette Brumback^b

This paper provides a brief overview to four major types of causal models for health-sciences research: Graphical models (causal diagrams), potential-outcome (counterfactual) models, sufficient-component cause models, and structural-equations models. The paper focuses on the logical connections among the different types of models and on the different strengths of each approach. Graphical models can illustrate qualitative population assumptions and sources of bias not easily seen with other approaches; sufficient-component cause models can illustrate specific hypotheses about mechanisms of action; and potential-outcome and structural-equations models provide a basis for quantitative analysis of effects. The different approaches provide complementary perspectives, and can be employed together to improve causal interpretations of conventional statistical results.

Keywords Bias, causal diagrams, causality, confounding, data analysis, direct effects, epidemiological methods, graphical models, inference, instrumental variables, risk analysis, sufficient-component cause models, structural equations

Accepted 28 March 2002

Following a long history of informal use in path analysis, causal diagrams (graphical causal models) saw an explosion of theoretical development during the 1990s,^{1–3} including elaboration of connections to other methods for causal modelling. The latter connections are especially valuable for those familiar with some but not all methods, as certain background assumptions and sources of bias are more easily seen with certain models, whereas practical statistical procedures may be more easily derived under other models. We provide here a brief overview of graphical causal models,^{1–6} the sufficient-component cause (SCC) models of Rothman,^{7,8} Ch. 2 the potential-outcome (counterfactual) models now popular in statistics, health, and social sciences,^{9–15} and the structural-equations models long established in social sciences.^{11–14} We focus on special insights facilitated by each approach, translations among the approaches, and the level of detail specified by each approach.

Graphical models

The following is a brief summary of terms and concepts of causal graph theory; see Greenland *et al.*⁴ and Robins⁵ for more detailed explanations. Figure 1 provides the graphs used for illustration below. An *arc* or *edge* is any line segment (with or without arrowheads) connecting two variables. If there is an

arrow from a variable X to another variable Y in a graph, X is called a *parent* of Y and Y is called a *child* of X. If a variable has an arrow into it (i.e. it has a parent in the graph) it is called *endogenous*; otherwise it is *exogenous*.

A *path* between two variables X and Y is a sequence of arcs connecting X and Y. A *back-door path* from X to Y is a path whose

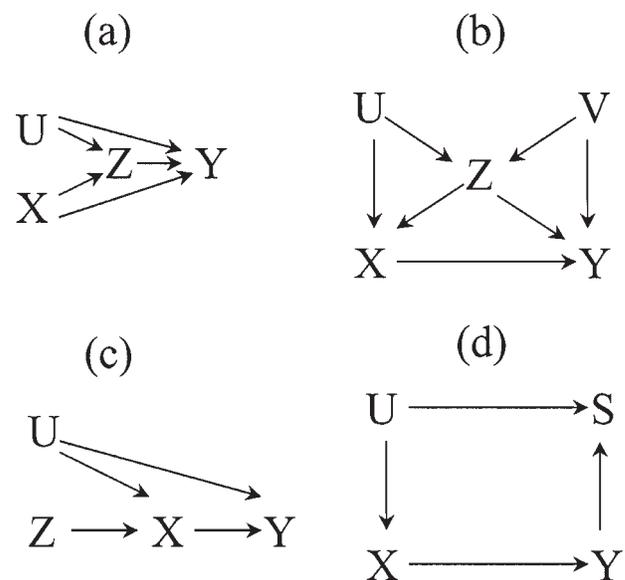


Figure 1 Four causal diagrams used in examples. In all four, X and Y are the exposure and outcome variables under study

^a Department of Epidemiology, UCLA School of Public Health, Department of Statistics, UCLA College of Letters and Science, 22333 Swenson Drive, Topanga, CA 90290–3434, USA. E-mail: lesdomes@ucla.edu

^b Department of Biostatistics, University of Washington, School of Public Health and Community Medicine, Seattle, WA 98195, USA.

first arc is an arrow pointing to X; there is no back-door path from X to Y in Figure 1a, whereas in Figure 1c the path X-U-Y is a back-door path from X to Y. A *blocked path* (or closed path) between X and Y is a path that passes from a parent to child and then back to another parent, i.e. there is a parent-child-parent sequence in the path; a path that has no such sequence is an *open path*.²⁻⁴ In Figure 1b, the paths X-U-Z-Y and X-Z-V-Y are open, but the path X-U-Z-V-Y is not because U-Z-V is a parent-child-parent sequence.

A *directed path* is a sequence of arrows such that the child in the sequence is the parent in the next step. If there is a directed path from X to Y, X is called an *ancestor* of Y and Y is called a *descendant* of X. A graph is *directed* if all the arcs in it are arrows; a graph is *acyclic* if no directed path forms a closed loop (equivalently, if no variable is both an ancestor and descendant of another). A graph that is both directed and acyclic is a DAG; each graph in Figure 1 is a DAG.

A graph is *causal* if every arrow represents the presence of an effect of the parent (causal) variable on the child (affected) variable. In a causal graph, a directed path represents a causal pathway, and an X-to-Y arrow represents a direct effect of X on Y within the graph (an effect not mediated through any other variable in the graph). Each graph in Figure 1 summarizes causal relations within a population of individuals, and each variable represents the states or events among individuals in that population. For example, if X is a treatment variable, then the value of X for an individual is the level of treatment received by the individual. Absence of a directed path from X to Y in the graph corresponds to the causal null hypothesis that no alteration of the distribution of X could change the distribution of Y.

The 'population' might contain just one individual, in which case the graph is a model for effects on that individual. Furthermore, the 'individuals' in the population need not be persons; they may be administrative entities, natural groupings, or any other unit of interest. For example, in a study of the effect of state helmet laws on riding-accident mortality Y among motorcyclists, the individual units could be states, X could be helmet-law status, and Z could be helmet-law enforcement levels. One could also draw an accident-level graph in which X could be helmet-law status in the accident's locale, Z indicates whether the motorcyclist was wearing a helmet, and Y indicates whether the motorcyclist was killed.

An important result from graph theory is that if one stratifies (conditions) on a descendant Z of two variables U and X, and U and X are independent in the total population, then we should expect U and X to be associated within at least one stratum of Z (exceptions to this rule involve somewhat contrived cancellations of effects).^{2,3 p. 17,4} To illustrate a consequence of this result, suppose in Figure 1a X represents a 6-month weight loss regimen that is randomly assigned within a cohort of cardiovascular patients, with X = 1 for regimen assigned and X = 0 for not assigned; Z represents a set of clinical CHD risk factors (serum lipids, blood pressure) measured at regimen completion; Y represents death within the year following completion; and U represents a set of unmeasured genes that affect death risk both directly and through the clinical factors Z. Although U affects Y, it is not a confounder of the X-Y association because it is independent of X.

A common approach to analysing effects of weight on health is to adjust for serum lipids and blood pressure. If weight affects

serum lipids and blood pressure, such adjustment cannot be justified as confounding control because it removes that part of the weight effect mediated through serum lipids and blood pressure.^{8 Ch. 4} It is often thought that such an analysis estimates the direct effect of weight, or of a weight-loss regimen. Using counterfactual models, however, it has been shown that this rationale fails if the intermediates were also affected by uncontrolled risk factors; it fails even if the treatment X is independent of the uncontrolled factors, so that there is no confounding of the crude X-Y association, as in Figure 1a.¹⁶ Graph theory shows this fact more simply: Because Z is a child of both U and X, one should expect U and X to be associated within at least one stratum of Z; consequently, within strata of Z, U becomes a confounder, even though it was not one to begin with.⁶ In general, one should expect control of an intermediate Z to generate confounding when Z and Y share causes other than X, as in Figure 1a; in such cases the association of Z with Y is confounded, and so the estimated indirect effect of X on Y being 'removed' by Z-adjustment is confounded.¹⁷

Figure 1b gives another example, which has a counter-intuitive quality and had to wait for graph theory for discovery. In this graph we ask, 'is it sufficient to stratify only on Z in order to unbiasedly estimate the effect of X on Y?' A common intuitive answer is 'Yes,' because physically preventing individual variation in Z would block the effects of U on Y and V on X and thus eliminate confounding by U and V (as well as confounding by Z). But in an observational study U and V would ordinarily be associated within some strata of Z, because they both affect Z. Within those strata, U would be associated with Y (through V) as well as with X, and V would be associated with X (through U) as well as with Y; consequently, both U and V would be confounders and one or the other would have to be controlled to remove the confounding.^{2,4}

One can recognize the insufficiency of controlling Z alone given Figure 1b in more traditional ways: The association of Z with Y given X is confounded by V; because adjustment for Z alone depends on this confounded association, one might conclude correctly that such adjustment could mislead, and that adjustment for V as well as Z would remedy the problem. But graphical theory also shows that adjustment for U rather than V would also suffice: because the V-X association produced by Z-adjustment is mediated entirely through U, U-adjustment eliminates confounding by V within Z strata.

The preceding examples illustrate how causal graphs supply simple visual methods to check for confounding and for sufficiency of confounder adjustment. Some basic results are: (1) an open back-door path from X to Y can produce an association between X and Y, even if X has no effect on Y, and so can produce confounding; (2) adjustment for certain variables can produce open back-door paths, and so produce confounding; (3) the X-Y association will be unconfounded if the only open paths from X to Y are directed paths from X to Y (so that the only sources of X-Y association are effects of X on Y).^{2,3 Ch. 3,4} These results lead to general criteria for identifying sets of variables sufficient for control of confounding given a graph.^{2,3 Ch. 3,4}

Potential-outcome (counterfactual) models

Graphs display broad qualitative assumptions about causal directions and independencies in a population. Although it is

surprising how much can be deduced from such assumptions,¹⁻⁶ the deductions are only qualitative (e.g. confounding present or absent in a particular stratification). Usually, however, more precise deductions are needed, and such deductions require a quantitative model that specifies in detail what would happen under alternative possible patterns of intervention or exposure. One class of quantitative models originating with Neyman and Fisher in the early 20th century⁹ are the *counterfactual* or *potential-outcome* models.⁸ Ch. 4,9-12,15 These models formalize notions of cause and effect found in much of philosophy and epidemiology,^{15,18,19} such as this passage from MacMahon and Pugh: '... an association may be classed presumptively as causal when it is believed that, had the cause [exposure] been altered, the effect [outcome] would have changed'.¹⁹ p. 12 A key feature of this description is its *counterfactual* element: It refers to what would have happened if, contrary to fact, the exposure had been something other than what it actually was.

Suppose we have a population of individual units under study (e.g. mice, people, counties) indexed by $i = 1, \dots, N$, a treatment or exposure variable X with $J + 1$ levels (or actions) x_0, x_1, \dots, x_J , and an outcome variable of interest Y (such as an indicator for 'death by age 70'). The standard potential-outcome model for a non-contagious outcome assumes that:

(a) Each individual could have received any one of the treatment levels; this rules out (for example) having men in the population for an analysis of hysterectomy effects.

(b) For each individual i and treatment level x_j , at the time of treatment assignment the outcome that individual i would have if the individual gets treatment level x_j exists, even if the individual does not in fact get x_j ; this value is called the *potential outcome* of individual i under treatment x_j .

The variable Y represents a generic variable for the actual outcome under the treatment actually given. Assumption (b) can be recast as stating that, for each individual i and each exposure level x_j , one can also define a potential-outcome (potential-response) variable Y_{ij} representing the outcome of the individual under that exposure. Thus, if Y is an indicator for 'death by age 70', Y_{ij} will be an indicator for 'death by age 70 of individual i if that individual is given treatment x_j '.

If individual i gets treatment x_j , Y_{ij} will equal the indicator for the actual outcome of individual i ; but otherwise it may be quite different from that actual outcome. Such a difference is taken as the effect of actual treatment relative to treatment x_j . More generally, the choice of treatment is said to have had no effect on Y for individual i if $Y_{ij} = Y_{ik}$ for every possible pair of treatment levels x_j and x_k ; otherwise, if $Y_{ij} \neq Y_{ik}$ for some pair of treatment levels x_j and x_k , treatment choice could have had an effect, or could have caused a change in the actual outcome of individual i (from Y_{ik} to Y_{ij}). Treatment choice is said to have had no effect on the population if it had no effect on any individual in the population.

In addition to (a) and (b), most applications also assume that the potential outcomes of each individual are independent of the treatments and outcomes of other individuals. This assumption is not always correct (e.g. in vaccine trials), but the model can be generalized to allow for violations.²⁰ A controversial aspect of assumption (b) is that it requires each potential outcome Y_{ij} remain a meaningful quantity even when individual i does not get treatment x_j . Even if one accepts this

idea, the only Y_{ij} that can be observed for individual i is the one corresponding to the treatment actually received by that individual; the remaining Y_{ij} can only be estimated, not observed. People routinely estimate such quantities in day-to-day life (e.g. 'if I had only bought Microsoft stock when it was first issued, my net worth would be millions of dollars'). The problems attributed to modelling such quantities (such as the need for untestable assumptions in estimating causal effects) are in reality unpleasant intrinsic problems of causal inference that are obscured by other approaches; we believe it is a virtue of the counterfactual approach that it makes such problems explicit.^{15,21}

Potential-outcome models are not inherently deterministic (as is often mistakenly claimed), because the potential outcomes (the Y_{ij}) may be parameters of probability distributions (e.g. expected age at death) rather than directly observable events (e.g. actual age at death).²¹ This flexibility can be seen in the probabilistic notations based on the 'set' and 'do' operators in Pearl,^{2,3} which can be used to represent effects in a single individual instead of a population. Furthermore, potential-outcome models are not limited to person-level analyses; for example, the 'individuals' in the model may be social units or aggregates (although the associations observed among these aggregates may be confounded by person-level effects).²²

One way of summarizing the scope of potential-outcome models is that they represent the limit of what one could learn about individual causes and effects from perfect crossover trials. For example, if X and Y represent completely reversible exposure and outcome variables (e.g. as might occur with X indicating a nasal irritant and Y a sneezing probability), we could estimate an individual's Y_{i1} and Y_{i0} (sneezing probabilities when irritant present and absent) through a series of trials on the individual that alternated $X = 1$ with $X = 0$, provided there were no carry-over effects or temporal variations in the sneezing responses (as represented by the potential outcomes). When such trials cannot be performed, as is usual in human studies, we could still estimate the population distribution of Y_{ij} (the outcome when $X = x_j$) by treating a random sample from that population with x_j . By repeating such experiments for various treatment levels (or by randomizing a random sample to different treatment levels) we can estimate how the population outcome distribution would vary with treatment distribution.^{9,11,15,21}

A practical aspect of potential-outcome models arising from assumption (b) is that any potential outcome Y_{ij} not observed (whether because treatment x_j was not given to i , or because of censoring) can be viewed as a quantity to be estimated or imputed from observed covariates and outcomes.^{9,23} This idea underlies most methods of model-based standardization of effect estimates,⁸ Ch. 21 and leads to numerous methods for confounder control based on the relation of actual treatment X to the potential outcomes predicted from various models.²³ Some effect measures do not require that assumptions (a) and (b) apply to all individuals in the study. For example, if unexposed ($X = 0$) individuals are used only to estimate the distribution of the Y_{i0} among the exposed ($X = 1$), as in many occupational studies, we need not assume that the unexposed could have been exposed or that Y_{i1} is meaningful for the unexposed.¹⁵

Multifactorial causation and the sufficient-component cause model

The graphical and potential-outcome models can be used to portray the presence, though not the mechanics, of causal interactions. Consider for example the synergism between phenylketonuria (PKU) ($X = 1$) and significant phenylalanine consumption (SPC) ($Z = 1$) in inducing brain damage ($Y = 1$): In some people, these two factors together are necessary and sufficient to produce damage.¹⁹ This synergism can be represented in basic graphs^{1–6} by including a variable XZ that indicates their joint presence ($XZ = 1$ if $X = Z = 1$ are both present, $XZ = 0$ otherwise), then drawing an arrow from XZ to Y . To represent the synergism in a potential-outcome model, we may define four damage indicators Y_{ixz} for each individual i ; the subscript x is 1 with PKU present, 0 with PKU absent, while z is 1 with SPC present, 0 with SPC absent. The synergism then corresponds to $Y_{i11} = 1$ but $Y_{i10} = Y_{i01} = Y_{i00} = 0$.^{8 Ch. 18}

Because potential outcomes are quantities specific to individuals in the modelled population, they provide more detail than arrows in graphs. For example, the individuals affected by X and those affected by Z may be one and the same, or may not overlap at all. The graph $X \rightarrow Y \leftarrow Z$ would hold if the population were composed entirely of individuals with $Y_{i11} = Y_{i10} = Y_{i01} = 1$ and $Y_{i00} = 0$; in this case, if everyone had their actual X and Z equal to 0, everyone would be affected by changes in X or changes in Z . But the same graph would hold if the population was half individuals with $Y_{i11} = Y_{i10} = 1$, $Y_{i01} = Y_{i00} = 0$ (individuals affected only by changes in X) and half individuals with $Y_{i11} = Y_{i01} = 1$, $Y_{i10} = Y_{i00} = 0$ (individuals affected only by changes in Z). Like the graph, the potential-outcome models can be extended to include effects of X on Z as well as the effects on Y ; doing so reveals many distinctions not captured by simply adding an arrow from X to Z in the graph.¹⁶ Such examples show that potential-outcome models are logically finer (distinguish more situations) than graphical models of the same variables; this fineness leads to greater notational complexity.

Consideration of causal mechanisms leads to models that are logically finer than either potential-outcome models or graphs. Best known among epidemiologists is Rothman's sufficient-component cause (SCC) model.^{7,8 Ch. 2} In this model, two factors are said to be causal *cofactors*, and have a (potential for) synergism, if they are components of the same causal mechanism; the presence of both cofactors is necessary for the mechanism to operate and so produce the outcome under study. This definition refers to mechanisms; thus, the basic units of analysis are the mechanisms that determine the potential outcomes of individuals, rather than individuals. Many different sets of mechanisms will lead to the same pattern of potential outcomes for an individual; hence, many different SCC models will lead to the same potential-outcome model.^{8 Ch. 18,25} As with potential-outcome models, however, SCC models are not inherently deterministic, because the component causes may be random events²⁴ and because the outcome affected by the completion of a sufficient cause may be a probability parameter rather than an observable event.

The SCC model employs a pie-chart representation of causal mechanisms, in which each slice represents a necessary component of the mechanism.⁷ To illustrate, suppose we are

considering mechanisms for angiosarcoma induction in just one individual i . Figure 2 gives an illustration of two distinct SCC models for the disease-causing mechanisms within this individual. The U in the figure represent sets of unmeasured cofactors that would be present regardless of this individual's X or Z status. Model (a) posits that there are two mechanisms that can lead to disease in this individual, neither of which involve synergism of X levels and Z levels, while model (b) posits three such mechanisms, all of which show synergism of X levels and Z levels. Nonetheless, under both models this individual will get the disease unless $X = 0$ and $Z = 0$; in other words, under either SCC model the individual's potential outcomes would be $Y_{i11} = Y_{i10} = Y_{i01} = 1$ and $Y_{i00} = 0$. Thus, even if we could conduct a perfect crossover trial on the individual and so observe the individual's outcome under all four X - Z combinations, we would still be unable to determine which SCC model was correct.

As this example and more realistic ones^{26–28} show, there are severe limits to the detail about causal mechanisms that can be distinguished using only ordinary ('black-box') randomized trials and epidemiological studies of exposure-disease relations.^{26–29} Although discrimination among mechanisms can be important,^{28,29} it will usually require direct observations of intermediate steps or of biomarkers for hypothesized mechanisms.

Structural-equations models

Informal use of graphs initially developed as an intuitive aid for structural-equations modelling (SEM), in which a web or network of causation is modelled by a system of equations and independence assumptions.^{3 Ch. 1,13} Each equation shows how an individual response (outcome, affected, dependent) variable changes as its direct (parent) causal variables change. Again, the 'individual' may be any unit of interest, such as a person or aggregate. In the system, a variable may appear in no more than one equation as a response variable, but may appear in any other equation as a causal variable. A variable appearing as a response in the system is said to be endogenous (within the system); otherwise it is exogenous.

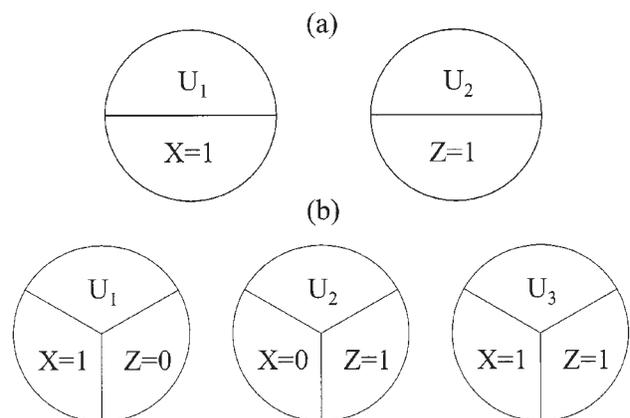


Figure 2 Two distinct sufficient-component cause (SCC) models for the set of mechanisms within an individual; each leads to the same potential-outcome model

A causal graph is a qualitative schematic for a class of structural-equations models. For example, Figure 1a is a schematic for the linear system

$$Z = \alpha_Z + \beta_{UZ}u + \beta_{XZ}x \quad (1a)$$

$$Y = \alpha_Y + \beta_{UY}u + \beta_{XY}x + \beta_{YZ}Z \quad (1b)$$

in which u , x , z are specific values of U , X , Z , α_Z and α_Y are unmeasured individual-specific (random) disturbances of Z and Y , and α_Z , α_Y , U and X are assumed to be jointly independent of one another in the study population. Figure 1a is also a schematic for the very different system

$$Z = \alpha_Z + \beta_{UZ}u + \beta_{XZ}x + \beta_{UXZ}ux \quad (2a)$$

$$\ln(Y) = \alpha_Y + \beta_{UY}u + \beta_{XY}x + \beta_{YZ}Z \quad (2b)$$

with α_Z , α_Y , U and X again assumed jointly independent. System 2 differs from system 1 in that a product term is added to the Z equation, and the Y equation is log-linear instead of linear. Nonetheless, both systems share the properties indicated by Figure 1a: U and X are the two exogenous variables (indicated by their lack of parents); U and X directly affect the two endogenous variables Z and Y , and Z directly affects Y (indicated by the arrows from U and X to Z and Y , and from Z to Y); and the exogenous variables and random disturbances are jointly independent of one another (indicated by the *absence* of connections among the variables other than the arrows just described).

Structural equations can be viewed as formulas for computing potential outcomes under various actions.^{2,3} Ch. 1 For example, if X represents a treatment regimen, equation 1a asserts that the *potential* value of Z for any individual in the study population will trace out a linear function of X as the individual's values of X changes but U remains constant: an individual's Z will change by β_{XZ} units if we increase X by one unit while U remains constant (because α_Z is constant for the individual). Equation 1a also asserts that Z will not vary with Y if U and X remain constant. It is such within-individual causal interpretations that distinguish structural equations from ordinary regression equations (which represent only *associations* of actual outcomes with actual values of the covariates as one moves across individuals).³ Ch. 5,8 Ch. 20 Structural-equations models (complete systems such as 1 and 2 above) combine potential-outcome models for the endogenous variables with independence assumptions about exogenous variables.

Structural equations with unknown parameters go beyond graphs by specifying the functional form of effects, but do not provide the exact values of effects; thus, they are algebraic but not fully quantified representations of causal relations. The equations can also be given a general non-parametric form that does not impose structure beyond that in the corresponding graph, and so is logically equivalent to that graph.^{2,3} Ch. 1 For example, Figure 1a corresponds to

$$Z = f_Z(u, x, \alpha_Z) \quad (3a)$$

$$Y = f_Y(u, x, z, \alpha_Y) \quad (3b)$$

with α_Z , α_Y , U , X assumed jointly independent. The functions f_Z and f_Y are left unspecified, although statistical analysis will

usually require some restrictions on f_Z and f_Y , such as smoothness of dose-response and additivity of effects on a particular scale. Equations 3a and 3b may be interpreted as alternative notations for the potential outcomes corresponding to Z and Y . For example, $f_Y(u, x, z, \alpha_Y)$ may be interpreted as the potential outcome y_{iuxz} , with the individual identifier i replaced by a 'random' source of inter-individual variation α_Y . Thus, non-parametric structural equation models provide a bridge between graphical and potential-outcome models.²

As with potential-outcome models, structural-equations models extend beyond deterministic outcomes, although the details of such extensions are rather technical. In the systems above, Z and Y may represent parameters of individual outcome distributions, rather than the observable outcome events. For example, Z and Y may represent expected values; the structural equations are then mixed models with random intercepts α_Z and α_Y . A common equivalent practice adds mean-zero 'random errors' ϵ_Z and ϵ_Y to the Z and Y equations; Z and Y then remain observable outcomes, but the random errors are not separable from α_Z and α_Y without repeated observations of all variables on each individual. It is also possible to treat the β coefficients as random.

Graphical versus algebraic representations

As an illustration of the differing insights obtained from graphical and algebraic representations of causation, Figure 1c diagrams a situation in which Z is an *instrumental variable* for estimating X effects: Z affects X , but is unassociated with the confounder U and is unassociated with Y except through X .³ Sec. 7.4.5 Such variables occur in randomized trials, in which Z is the assigned (intended) treatment. Many patients do not fully comply, and instead take (or receive) a different level of treatment, X ; this received-treatment variable is affected by unmeasured factors U that are also risk factors (or close correlates of risk factors) for the outcome under study. Standard intent-to-treat analyses examine only the Z association with Y and so are estimating the effect of treatment *assignment*, rather than a physiologic effect of received treatment X . Can we also estimate the latter effect? The answer is yes, provided we can make further (not necessarily unique) quantitative assumptions. The graph makes clear that we should not expect the crude X - Y association to equal the X - Y effect, because of confounding by U . The graph also shows, however, that there is no confounding of the Z effects on X or Y (as would be expected if Z was randomized); hence the crude Z - X and Z - Y associations will equal the Z - X and Z - Y effects. These facts alone can allow one to put bounds on the X - Y effect,³ Sec. 8.4 although one or both bounds may be beyond any plausible range for the effect.³⁰

Suppose we go beyond Figure 1c by assuming the linear structural relations

$$X = \alpha_X + \beta_{UX}u + \beta_{ZX}Z \quad (4a)$$

$$Y = \alpha_Y + \beta_{UY}u + \beta_{XY}X \quad (4b)$$

with α_X , α_Y , U , Z jointly independent. As noted long ago by economists,³¹ this model would allow us to unbiasedly estimate β_{XY} from the simple regressions of X on Z and Y on Z . First, because α_X , U , and Z are independent, there would be no

confounding of the simple β_{ZX} estimate obtained from regressing X on Z alone. Second, we can substitute 4a into 4b to get

$$\begin{aligned} Y &= \alpha_Y + \beta_{UY}u + \beta_{XY}(\alpha_X + \beta_{UX}u + \beta_{ZX}Z) \\ &= (\alpha_Y + \alpha_X\beta_{XY}) + (\beta_{UY} + \beta_{UX}\beta_{XY})u + \beta_{ZX}\beta_{XY}Z \\ &= \delta_Y + \delta_{UY}u + \delta_{ZY}Z, \end{aligned} \quad (5)$$

where $\delta_Y \equiv \alpha_Y + \alpha_X\beta_{XY}$, $\delta_{UY} \equiv \beta_{UY} + \beta_{UX}\beta_{XY}$, and $\delta_{ZY} \equiv \beta_{ZX}\beta_{XY}$. Because of the independence assumptions, there would be no confounding of the simple δ_{ZY} estimate obtained from regressing Y on Z alone; therefore, the ratio of the simple δ_{ZY} and β_{ZX} estimates will consistently estimate

$$\delta_{ZY}/\beta_{ZX} = \beta_{ZX}\beta_{XY}/\beta_{ZX} = \beta_{XY}, \quad (6)$$

which is just the effect of X on Y in system 4. This ratio is an example of an *instrumental-variables estimate* of effect,³ Sec. 3.5,30–32 one can also easily derive this estimate for binary X, Y, and Z by specifying potential outcomes directly.^{30,32} In either approach, it is important to remember that equation (6) is based on the linearity assumptions seen in system 4, as well as on the directional assumptions in Figure 1c.

For instrumental variables, algebraic modelling led to discovery of assumptions (plausible in some settings) that are sufficient for estimating the effects of interest from the given data. Nonetheless, by focussing our attention on basic qualitative relations, graphs can help identify fallacies in causal inference. Some examples were given in our discussion of Figures 1a and 1b; as another example, some epidemiologists still believe (mistakenly) that an extraneous factor cannot induce selection bias unless it is a risk factor for disease. Consider a case-control study of magnetic-field exposure X and childhood leukaemia Y, with U representing socioeconomic factors and S selection. It has been argued (though disputed) that socioeconomic factors have little or no effect on childhood-leukaemia risk (as opposed to diagnosis or mortality); there is evidence, however, that those factors are associated with magnetic fields and with participation.^{33,34} Because of the case-control design, leukaemia is also strongly associated with selection. Figure 1d summarizes this background. It shows that S is a descendant of both U and Y; hence, because the study data must be limited to those selected (the S = 1 stratum), we should expect U and Y to be associated in those data even if U has no effect on Y. Consequently, U would have to be controlled in order to ensure an unbiased estimate of the X-Y effect. Such control could not be accomplished if U were unmeasured or poorly measured. (Note however that if X itself affected selection, there would be no way to remove the resulting selection bias through control of a covariate.)

Discussion

What population should be modelled?

When using models in data analysis, it is essential to consider the distribution of exposure and confounders in the combined study population of all treatment (or exposure) groups that are under comparison, not in some specific target group of policy interest. Furthermore, in a population-based case-control study

this population will be the source population of cases and controls, not just the subjects selected into the study.⁸ Ch. 7 For example, a study of vinyl chloride effects may have as its target only workers actually exposed; nonetheless, to evaluate confounding one needs to include the unexposed group (as well as exposed group) in the population being modelled. Even though the target comprises only those exposed (X = 1), an unexposed population is needed for comparison, and whether or not an extraneous factor (indicated by U = 1) is a confounder depends on whether or not the factor is associated with the exposure in the entire (exposed plus unexposed) study population.⁸ Ch. 8 This pivotal U-X association can only be represented in a model for relations in the entire population (among the exposed, X is always 1 and so cannot be associated with anything).

What is a causal variable?

A controversial issue in all theories of causation is whether a variable must be manipulable to be considered potentially causal. For modelling purposes, some authors would restrict the label 'causal' to variables that represent interventions or actions,³⁵ or at most allow only mutable variables (those susceptible to intervention) as potentially causal.³ Such restrictions exclude as causal those variables regarded as immutable or defining characteristics of individuals, such as the birthdate and genetic sex of persons, but allow as causal such variables as perceived age and sex. Even when technology advances enough to allow alteration of a previously immutable characteristic (e.g. through genetic engineering), some authors would only label as 'causal' the intervention that alters the characteristic.³⁵

In potential-outcome models, the levels of immutable variables may be represented by strata (i.e. subpopulations) but not by interventions (i.e. not by x_j). In graphical and structural models, immutable variables may appear as exogenous variables, and so are not distinguished from manipulable exogenous variables. This practice is more in accord with ordinary usage of 'causal'; it is useful because all the graphical rules for assessing bias sources and covariate control continue to apply when including immutable variables.³ Ch. 3 The distinction between mutable and immutable variables remains important, however, as it leads to refinement of vague concepts like 'race' into multiple variables that have very different implications for health outcomes (e.g. mutable variables such as ethnic identification, and immutable variables such as ancestry).³⁶

A more severe problem arises when variables that are not interventions are treated as interventions for planning purposes.³⁶ A common example is estimation of 'the effect' of eliminating a disease (e.g. lung cancer) on life expectancy. This effect is quite dependent on how the disease is eliminated; for example, if it is eliminated by chemoprevention or vaccination, there may be occasional fatal side effects, or there may be causal or preventive effects on other potentially fatal diseases. Careful consideration of the ambiguities inherent in 'disease elimination' should lead instead to estimation of the effect of specific interventions designed to reduce or eliminate the disease burden.³⁶

Conclusions

Of the four causal modelling methods reviewed here, SCC models (the only ones originating in epidemiology) stand apart

in requiring specification of mechanisms within the individual units under study. There are rarely data to support such detailed specification, which may explain why SCC models have seen little use beyond teaching examples. Structural equations have seen extensive analytic application (especially in the social sciences^{10,12,13,31}), and potential-outcome models have been used to derive permutation tests for randomized trials for 80 years.⁹ Nonetheless, in epidemiology these models remain confined largely to the conceptual teaching realm (to the extent that they appear at all).^{8,37} This confinement may be partly due to their absence from current training: unfamiliar techniques are rarely used. Furthermore, the most recent innovations based on potential outcomes (g-estimation^{38,39} and marginal structural modelling⁴⁰) are designed for longitudinal data on time-varying exposures and confounders, which precludes their use in many if not most studies; the techniques also require special programming.

Due to their qualitative form, graphical models have not led to as many analytic techniques as have algebraic models. On the other hand, they can be easily applied in any study to display assumptions of causal analyses, and to check whether covariates or sets of covariates are insufficient, excessive, or inappropriate to control given those assumptions.^{1–6,14,17} When those assumptions are in doubt, one can still formulate a series of plausible graphs and conduct a corresponding series of analyses.⁴¹ Constructing graphs to accompany conventional statistical analyses of effects can at least help avoid or spot common mistakes, such as control of intermediates as if they were confounders.^{6,14,17}

Acknowledgements

The authors would like to thank Charles Poole, Judea Pearl, Katherine Hoggatt, and a referee for helpful comments on this paper.

KEY MESSAGES

- There are now at least four major classes of causal models in the health-sciences literature: Causal diagrams (graphical causal models), potential-outcome models, structural-equations models, and sufficient-component cause models.
- Causal diagrams can provide an easily understood depiction of qualitative assumptions behind a causal analysis, while potential-outcome and structural-equations models can depict more detailed quantitative assumptions about responses of units comprising the study population.
- Sufficient-component cause models differ from the other models in that they depict more elaborate qualitative assumptions about causal mechanisms within population units.

References

- 1 Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. New York: Springer, 1993.
- 2 Pearl J. Causal diagrams for empirical research (with discussion). *Biometrika* 1995;**82**:669–710.
- 3 Pearl J. *Causality*. New York: Cambridge, 2000.
- 4 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
- 5 Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;**12**:313–20.
- 6 Hernán MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation. *Am J Epidemiol* 2002;**155**:176–84.
- 7 Rothman KJ. Causes. *Am J Epidemiol* 1976;**104**:587–92.
- 8 Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd Edn. Philadelphia: Lippincott, 1998.
- 9 Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Ann Rev Public Health* 2000;**21**:121–45.
- 10 Winship C, Morgan SL. Estimation of causal effects from observational data. *Ann Rev Sociol* 1999;**25**:659–706.
- 11 Greenland S. Causal analysis in the health sciences. *J Am Statist Assoc* 2000;**95**:286–89.
- 12 Sobel M. Causal inference in the social sciences. *J Am Statist Assoc* 2000;**95**:647–51.
- 13 Heckman JJ, Vytlacil E. Econometric evaluations of social programs. In: Leamer E, Heckman JJ (eds). *Handbook of Econometrics*, Vol. 6. New York: Elsevier, 2003 (in press).
- 14 Kaufman JS, Kaufman S. Assessment of structured socioeconomic effects on health. *Epidemiology* 2001;**12**:157–67.
- 15 Maldonado G, Greenland S. Estimating causal effects (with discussion). *Int J Epidemiol* 2002;**31**:421–38.
- 16 Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3**:143–55.
- 17 Cole S, Hernán M. Fallibility in estimating direct effects. *Int J Epidemiol* 2002;**31**:163–65.
- 18 Levin ML. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* 1953;**9**:531–41.
- 19 MacMahon B, Pugh TF. Causes and entities of disease. In: Clark DW, MacMahon B (eds). *Preventive Medicine*. Boston: Little Brown, 1967, pp. 11–18.
- 20 Halloran ME, Struchiner CJ. Causal inference for infectious diseases. *Epidemiology* 1995;**6**:145–51.
- 21 Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14**:29–46.
- 22 Greenland S. Ecologic versus individual-level sources of confounding in ecologic estimates of contextual health effects. *Int J Epidemiol* 2001;**30**:1343–50.
- 23 Robins JM. Causal inference from complex longitudinal data. In: Berkane M (ed.). *Latent Variable Modeling with Applications to Causality*. New York: Springer, 1997, pp. 69–117.
- 24 Poole C. Positized epidemiology and the model of sufficient and component causes. *Int J Epidemiol* 2001;**30**:707–09.
- 25 Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ Health* 1988;**14**:125–29.
- 26 Siemiatycki J, Thomas DC. Biological models and statistical interactions. *Int J Epidemiol* 1981;**10**:383–87.

- ²⁷Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;**44**:221–32.
- ²⁸Beyea J, Greenland S. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Physics* 1999;**76**:269–74.
- ²⁹Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. *Jurimetrics* 2000;**40**:321–40.
- ³⁰Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Statist Assoc* 1996;**91**:444–72.
- ³¹Bowden RJ, Turkington DA. *Instrumental Variables*. Cambridge: Cambridge University Press, 1984.
- ³²Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29**:722–29.
- ³³Bracken M, Belanger K, Hellenbrand K *et al*. Correlates of residential wiring code used in studies of health effects of residential electromagnetic fields. *Am J Epidemiol* 1998;**148**:467–74.
- ³⁴Hatch EE, Kleinerman RA, Linet MS *et al*. Do confounding or selection factors of residential wiring codes and magnetic fields distort findings of electromagnetic field studies? *Epidemiology* 2000;**11**:189–98.
- ³⁵Holland PW. Statistics and causal inference (with discussion). *J Am Statist Assoc* 1986;**81**:945–60.
- ³⁶Greenland S. Causality theory for policy uses of epidemiological measures. In: Murray CJL, Mathers C, Salomon J, Lopez AD, Lozano R (eds). *Summary Measures of Population Health*. Cambridge, MA: Harvard, 2002, Ch. 6.2.
- ³⁷Newman SC. *Biostatistical Methods in Epidemiology*. New York: Wiley, 2001.
- ³⁸Witteaman J, D’Agostino RB, Stijnen T *et al*. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Study. *Am J Epidemiol* 1998;**148**:390–401.
- ³⁹Joffe MM, Hoover DR, Jacobson LP *et al*. Estimating the effect of zidovudine on Kaposi’s sarcoma from observational data using a rank-preserving structural failure-time model. *Stat Med* 1998;**17**:1073–102.
- ⁴⁰Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effects on zidovudine on the survival of HIV-positive men. *Epidemiology* 2000;**11**:561–70.
- ⁴¹Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;**9**:361–67.