

2B

Semantics of causal DAG models and the identification of direct and indirect effects

James M. Robins

Harvard School of Public Health, USA

Directed acyclic graphs (DAGs) are commonly used to represent causal models. The article by Dawid posits a causal model that is closely related to the model of Spirtes *et al.* (1993) and the model of Pearl (1993b). In this discussion I will compare and contrast the semantics of DAGs representing the Spirtes *et al.* model with that of DAGs representing the non-parametric structural equation (NPSE) model of Pearl (1995a) and the finest fully randomized causally interpreted structured tree graph (FRCISTG) model of Robins (1986). This discussion will be more philosophical than other contributions to this volume for the following reason: the major controversies in this field are often focused upon the causal rather than the statistical interpretation of various analytic procedures. For example, the finest FRCISTG and NPSE models both assume the existence of counterfactual variables; Dawid denies their existence, and the Spirtes *et al.* (1993) model is ‘agnostic’. To give the flavour of the issues involved I will review in detail one controversy. The controversy concerns the question of whether, when, and how the direct effects of a treatment on an outcome can be separated by means of statistical analysis from the treatment’s indirect effects. The discussion is organized as follows. In Section 1, I define the causal models that are to be compared. In Section 2, I collect mathematical results on the identification of direct and indirect effects. In Section 3, I discuss the substantive implications of these results.

1 Causal models and their DAG representation

We are given a DAG G with a vertex set of random variables $V = (V_1, \dots, V_M)$ with density $f_V(v)$ ordered so that V_j is not a descendant of V_m for $m > j$. We shall use the following notational conventions. For any random variable Z , we let a calligraphic \mathcal{Z} denote the support (i.e. the set of possible realizations z) of Z . For any z_1, \dots, z_m , define $\bar{z}_m = (z_1, \dots, z_m)$. By convention $\bar{z}_0 \equiv z_0 \equiv 0$. Let X denote any subset of V and let x be a realization of X . Both the finest FRCISTG and NPSE causal models assume the existence of the counterfactual random variable $V_m(x)$ encoding the value the variable V_m would have if, possibly contrary to fact, X were set to x , where $V_m(x)$ is assumed to be well defined in the sense that there is reasonable agreement as to the hypothetical intervention (i.e. closest possible world) that sets X to x (Robins and Greenland 2000).

Finest FRCISTG causal model A finest FRCISTG model assumes (i) all one-step-ahead counterfactuals $V_m(\bar{v}_{m-1})$ exist; (ii) $V_m(\bar{v}_{m-1}) \equiv V_m(pa_m)$ is a function of \bar{v}_{m-1} only through the values pa_m of V_m 's parents on G ; (iii) both the observed variables V_m and the counterfactuals $V_m(x)$ for any $X \subset V$ are obtained recursively from the $V_m(\bar{v}_{m-1})$, e.g. $V_3 = V_3\{V_1, V_2(V_1)\}$ and $V_3(v_1) = V_3\{v_1, V_2(v_1)\}$; and (iv)

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \perp\!\!\!\perp V_m \mid \bar{V}_{m-1} = \bar{v}_{m-1},$$

$$\text{for all } m \text{ and all } \bar{v}_{M-1} \in \bar{V}_{M-1}, \quad (1.1)$$

where \bar{v}_k is a subvector of \bar{v}_{M-1} for $k < M-1$.

NPSE causal model A NPSE model assumes that there exists mutually independent random variables U_m and deterministic unknown functions f_m such that the counterfactual $V_m(\bar{v}_{m-1}) \equiv V_m(pa_m)$ is given by $f_m(pa_m, U_m)$ and both the observed variables V_m and the counterfactuals $V_m(x)$ for any $X \subset V$ are obtained recursively from the $V_m(\bar{v}_{m-1})$ as above.

Under a NPSE causal model

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \perp\!\!\!\perp V_m(\bar{v}_{m-1}^{**}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}^*$$

$$\text{for all } m, \text{ all } \bar{v}_{M-1} \in \bar{V}_{M-1}, \text{ and all } \bar{v}_{m-1}^{**}, \bar{v}_{m-1}^* \in \bar{V}_{m-1}. \quad (1.2)$$

Thus a NPSE model is a finest FRCISTG but the converse is false, because a FRCISTG assumes independence of $\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\}$ and $V_m(\bar{v}_{m-1}^{**})$ given $V_{m-1} = \bar{v}_{m-1}^*$ only when $\bar{v}_{m-1}^{**} = \bar{v}_{m-1}^* = \bar{v}_{m-1}$.

In my 1995 *Biometrika* comment on Pearl (1995a), I proved the following:

Lemma 1.1 *If a DAG G represents a FRCISTG, then the density $f_V(V)$ of the observables V satisfies the Markov factorization*

$$f_V(v) = \prod_{j=1}^M f(v_j \mid pa_j). \quad (1.3)$$

Remark In my 1995 *Biometrika* comment, I incorrectly claimed in my Lemma 1 that a NPSE model and a finest FRCISTG were equivalent. Butch Tsiatis pointed out to me that I had failed to note that an NPSE model satisfied the stronger assumption of (1.2).

Before defining the agnostic causal model of Spirtes *et al.* (1993) we need to discuss intervention distributions and the g-computation algorithm functional.

Intervention distributions on FRCISTGs Suppose we are given a set of variables $X = \{X_1, \dots, X_k\} \subset V$ and an intervention DAG G^- that agrees with DAG G except the parents PA_m^- of $X_m \in X$ may differ from the parents of X_m on

G . A non-random G^- -specific treatment regime g^- is a collection of functions $g^- = \{g_1^-, \dots, g_k^-; g_m^-: \mathcal{P}A_m^- \rightarrow \mathcal{X}_m\}$ that gives the value $g_m^-(pa_m^-)$ that we will set X_m to when PA_m^- takes the value pa_m^- . When for each m , X_m has no parents on G^- , so that $g_m^-(pa_m^-)$ is a constant, say x_m^* , we say regime g^- is non-dynamic and write $g^- = x^* = \{x_1^*, \dots, x_k^*\}$. Otherwise, g^- is dynamic. The counterfactual random variable $V_j(g^-)$ associated with regime g^- is recursively defined as follows: (i) when $V_j \in V \setminus X$, $V_j(g^-)$ is the one-step-ahead counterfactual $V_j(\bar{v}_{j-1})$ evaluated at $\bar{v}_{j-1} = \bar{V}_{j-1}(g^-)$ and (ii) when $V_j = X_m \in X$, $V_j(g^-)$ is $g_m^-(pa_m^-)$ with pa_m^- equal to the counterfactual $PA_m^-(g^-)$.

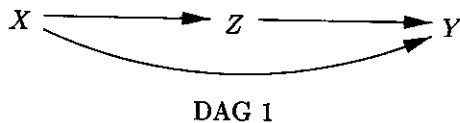
Lemma 1.2 (Robins 1986) *If DAG G represents a FRCISTG, then for any set of variables $X \subset V$, and associated intervention DAG G^- , and any treatment regime g^- , the (intervention) density $f_{V(g^-)}(v)$ of the counterfactual $V(g^-)$ is a functional of the density $f_V(v)$ of V and thus is non-parametrically identified from data V . This functional, which I have referred to as the g^- -computation algorithm functional or density (hereafter g^- -functional or density), is the density $f_{g^-}(v)$ obtained by modifying the product on the right-hand side of (1.3) as follows: if $V_j = X_m \in X$, remove the term $f(v_j | pa_j)$ from the product and set v_j to the value $g_m^-(pa_m^-)$ elsewhere in (1.3).*

Agnostic causal model The agnostic causal model of Spirtes *et al.* (1993) effectively assumes that the joint distribution of V factors as in (1.3) and that the joint density of V under the regime g^- on a graph G^- is given by the g^- -functional $f_{g^-}(v)$. Although this agnostic model assumes that the density of V under the intervention g^- is well defined, the model makes no reference to counterfactual variables and is agnostic as to their existence. In his article, Phil Dawid embraces a restricted version of the agnostic model in which only a subset of the variables in V can be manipulated (i.e. set) which he refers to as the decision variables. This model is closely related to the causal model discussed by Heckerman and Shachter (1995). The randomized causally-interpreted structured tree graph (RCISTG) of Robins (1987a,b) likewise restricts the set of variables in V that can be manipulated. The relationship of a RCISTG model to an FRCISTG model is analogous to that of Dawid's restricted agnostic model to the agnostic model. In his article Dawid was interested in the marginal intervention distribution $f_{g^-}(y) = \int \dots \int f_{g^-}(v) d\mu(y^c)$ of a subset Y of the variables in $V = (Y, Y^c)$, say, when data are obtained only on some subset V^* of the variables in V . In this case one wishes to know whether the intervention distribution $f_{g^-}(y)$ of Y is identified from (i.e. is a functional of) the marginal distribution $f_{V^*}(v^*)$ of V^* and, if not, to set bounds on $f_{g^-}(y)$. Sufficient conditions for identification have been derived by Galles and Pearl (1995) for univariate (i.e. time-independent) interventions, Pearl and Robins (1995) for non-dynamic regimes, and Robins (1997) for dynamic regimes. We refer the reader to the above references for additional discussion.

2 Direct and indirect effects

Define a causal DAG model to be a manipulative causal DAG model if the only causal effects that are non-parametrically identified from the joint distribution of the variables on the DAG are those that could in principle be checked by manipulation of (equivalently, experimental intervention on or setting of) the DAG variables. That is, a manipulative causal model is one in which all the causal predictions of the model can in principle be checked (i.e. tested) by experimental intervention. The finest FRCISTG model is a manipulative model because the causal parameters that are non-parametrically identified from data on V are all functions of the counterfactual intervention densities $f_{V(g^-)}(v)$. Specifically, suppose we measure data on all the variables V on G in a large study population so that we can regard $f_V(v)$ as known. Then to check our finest FRCISTG causal model, we could take an as-yet-untreated population exchangeable with the study population and intervene by forcing them to follow some regime g^- , allowing us to empirically estimate the intervention distribution $f_{V(g^-)}(v)$. If, for some regime g^- associated with a graph G^- , the g^- -functional $f_{g^-}(v)$ differs from the intervention distribution $f_{V(g^-)}(v)$, then we can conclude that our causal model is false, as would occur if there were a common cause of two variables in V that was not itself included in V . This argument also implies that the agnostic causal model is a manipulative model. Of course in practice such intervention tests may be impossible to carry out for logistical reasons (e.g. some variables V_m cannot be measured or there is no untreated population that one regards as exchangeable with the study population) or for ethical reasons.

If, however, a causal model is non-manipulative and thus non-parametrically identifies causal effects that do not correspond to the effect of an experimental intervention, then there is no way, even in principle, that one could check the correctness of all the model predictions. Robins (1986) imposed the independence assumption (1.1) precisely because (1.1) is the independence assumption that identifies all manipulative effects $f_{V(g^-)}(v)$ without identifying any non-manipulative effects. In the course of the following discussion of direct and indirect effects, we show the NPSE model is a non-manipulative model, because it implies the stronger independence assumption (1.2).



Robins and Greenland (1992) (hereafter R&G) define the pure direct effect (PDE) of a (dichotomous) exposure X on Y not acting through the intermediate variable Z to be the mean of Y under exposure to X had, contrary to fact, X 's effect on the intermediate Z been blocked (that is, had Z remained at its value under non-exposure thereby eliminating all indirect effects) minus the mean of

Y under non-exposure to X . That is, under a NPSE or FRCISTG model,

$$\begin{aligned} \text{PDE} &= \mathbf{E}[Y \{x = 1, Z(x = 0)\}] - \mathbf{E}[Y(x = 0)] \\ &= \mathbf{E}[Y \{x = 1, Z(x = 0)\}] - \mathbf{E}[Y(x = 0, Z(x = 0))], \end{aligned}$$

since $\mathbf{E}[Y(x = 0)] = \mathbf{E}[Y(x = 0, Z(x = 0))]$. Here $Y(x, z)$ is the counterfactual value of Y with (X, Z) set to (x, z) and $Z(x)$ is the counterfactual value of Z when X is set to x . The total indirect effect (TIE) of a (dichotomous) exposure X on Y is the total effect of X on Y minus the PDE. The motivation underlying this definition is that any effect of X on Y that is not purely direct must have an indirect contribution. Thus,

$$\begin{aligned} \text{TIE} &= \mathbf{E}[Y(x = 1) - Y(x = 1, Z(x = 0))] \\ &= \mathbf{E}[Y(x = 1, Z(x = 1))] - \mathbf{E}[Y \{x = 1, Z(x = 0)\}], \end{aligned}$$

since $\mathbf{E}[Y(x = 1)] - \mathbf{E}[Y(x = 0)]$ is by definition the total (equivalently, net or overall) exposure effect.

Similarly, the pure indirect effect (PIE) of X on Y through an intermediate variable Z is defined to be the mean of Y under non-exposure to X but with Z set to its exposed value minus the mean of Y under non-exposure to X . That is, under a NPSE or FRCISTG model,

$$\text{PIE} = \mathbf{E}[Y(x = 0, Z(x = 1))] - \mathbf{E}[Y(x = 0)]. \quad (2.1)$$

In this contrast the only effect of X on Y is indirect in that the effect is relayed through X 's effect on Z . The total direct effect (TDE) of X on Y not through an intermediate variable Z is the total effect of X on Y minus the PIE. Thus,

$$\text{TDE} = \mathbf{E}[Y(x = 1)] - \mathbf{E}[Y(x = 0, Z(x = 1))].$$

Pearl (2001) adopted our definitions but changed nomenclature. He refers to pure direct and indirect effects as natural direct and indirect effects. Under the agnostic causal model, the concept of the total and pure, indirect and direct effects is not defined since the counterfactuals $\mathbf{E}[Y \{x = 1, Z(x = 0)\}]$ and $\mathbf{E}[Y(x = 0, Z(x = 1))]$ are not assumed to exist.

The direct effect of X when Z is set to z (i.e. $\mathbf{E}[Y(1, z)] - \mathbf{E}[Y(0, z)]$) is identified from $f_Y(v)$ by the g-formula under all three models. But, in general, the contrast $\mathbf{E}[Y(1, z)] - \mathbf{E}[Y(0, z)]$ differs from both the PDE and TDE contrasts and differs depending on whether z is set to 1 or to 0. Indeed, since the intervention mean $\mathbf{E}[Y(x)]$ is identified under any of the three causal models, determining whether the TIE, PIE, TDE, and PDE are identified is equivalent to determining whether $\mathbf{E}[Y(x, Z(x^*))]$, $x \neq x^*$, is identified. Below we will prove that $\mathbf{E}[Y(x, Z(x^*))]$ is not a manipulative effect and thus is not identified under a finest FRCISTG model. However, Pearl (2001) proved that under a NPSE model $\mathbf{E}[Y(x, Z(x^*))]$ is identified for certain DAGs. As an example

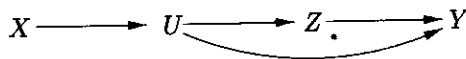
Consider a DAG, such as DAG 1 or DAG 2, where X has no parents and Z is a non-descendant of Y . Pearl showed that if

$$Y(x, z) \perp\!\!\!\perp Z(x^*) \mid X \quad \text{for all } z, \quad (2.2)$$

then under either a FRCISTG or NPSE model

$$E[Y(x, Z(x^*))] = \int E[Y(x, z)] dF_{Z(x^*)}(z). \quad (2.3)$$

Equation (2.3) is non-parametrically identified from $f_V(v)$ under all three causal models since the intervention parameters $E[Y(x, z)]$ and $f_{Z(x^*)}(z)$ are identified by the g -formula. However, no finest FRCISTG model implies (2.2). On the other hand, it follows from (1.2) that (2.2) holds for the NPSE model represented by DAG 1. Note that (2.2) will not hold for the NPSE model represented by DAG 2. Indeed $E[Y(x, Z(x^*))]$ will never be identified from $f_V(v)$ for an NPSE model represented by any DAG which contains a descendant C of X that is an ancestor of both Z and Y . In summary, under a NPSE causal model, PIE and PDE would be non-parametrically identified based on DAG 1 but not on DAG 2.



DAG 2

A non-manipulative model We now turn to the question of whether $E[Y(x, Z(x^*))]$ can be identified by manipulating (i.e. setting) the variables on G . Now, as noted by R&G, we could identify $E[Y(x, Z(x^*))]$ if we could manipulate X to x^* , observe $Z(x^*)$, then ‘return each subject to their pre-intervention state’, manipulate X to x and Z to $Z(x^*)$, and finally observe $Y(x, Z(x^*))$. However, such an intervention strategy usually will not exist because we cannot ‘return each subject to their pre-intervention state’ by any conceivable real-world intervention (as, for example, if the outcome Y were death). As a result, because we cannot observe the same subject under both $X = x$ and $X = x^*$, we are unable to directly observe the joint distribution of $Z(x)$ and $Z(x^*)$. It follows that we cannot identify $E[Y(x, Z(x^*))]$ by any manipulation of the variables on G owing to the impossibility of differentiating exposed ($x = 1$) subjects, whose value of Z is attributable to X (i.e. $Z(x) \neq Z(x^*)$) from those whose value of Z is not ($Z(x) = Z(x^*)$). We will thus refer to $E[Y(x, Z(x^*))]$ and the pure and total, direct and indirect effect contrasts as non-manipulative parameters.

From the above, we conclude that the NPSE causal model is a non-manipulative model. One could argue that manipulative models are preferable because the predictions of non-manipulative models are not, even in principle, testable by experiment. Here are some counter-arguments defending the use of non-manipulative models.

First, the assumptions required to identify $E[Y(x, Z(x^*))]$ are analogous to the assumptions necessary to identify manipulative causal effects such as the total effect $E[Y(x)] - E[Y(x^*)]$ from observational (i.e. non-randomized) data, and we do not wish to argue against using observational data to estimate effects of exposures that cannot be tested experimentally for ethical or logistical reasons. The following sentence makes the analogy. Because we cannot observe the same subject under both $X = x$ and $X = x^*$, we can generally only (i) identify $E[Y(x)] - E[Y(x^*)]$ if we assume that within levels of baseline variables $Y(j) \perp\!\!\!\perp X$, $j = 0, 1$, and (ii) identify $E[Y(x, Z(x^*))]$ if we assume (2.2) holds within levels of baseline variables. Secondly, the fact that our observational study estimate of $E[Y(x)] - E[Y(x^*)]$ could, in principle, be checked by experiment is of no import when in fact the check cannot be carried out for either ethical or logistical reasons. Third, even when an experiment can be conducted, if one is uncertain that the available experimental subjects are exchangeable with the observational study subjects, an observational estimate of $E[Y(x)] - E[Y(x^*)]$ cannot be checked, as it is possible that any difference between observational and experimental estimates is wholly due to a lack of exchangeability. Fourth, suppose one would assume DAG 1 is an FRCISTG so that

$$Y(x, z) \perp\!\!\!\perp Z(x) \mid X = x \quad \text{for all } z, x. \quad (2.4)$$

It is hard to construct realistic (as opposed to mathematical) scenarios in which one would accept (2.4) as true but not (2.2), as it is unlikely one would accept either (2.4) or (2.2) as true unless one believed that $Z = Z(X)$ was effectively randomly assigned by nature within levels of X , in which case both (2.2) and (2.4) would be true.

An alternative identifying assumption Even if one is convinced by the above four arguments for using an NPSE model, I suspect that, in practice, this model could rarely be used to identify the non-manipulative parameters corresponding to pure and total, direct and indirect effects as it would be unusual to have sufficient prior causal knowledge to impose the identifying assumption that there is no variable, say U , that is both affected by X and is a common cause of Z and Y (i.e. there is no descendant U of X that is an ancestor of both Z and Y). (Note that if such a U exists, then it must be included on the DAG G in order for G to represent any of our three causal models.) Thus, one might wish to consider alternative identifying assumptions such as the following no-interaction assumption.

No-interaction assumption $Y(x, z) - Y(x^*, z)$ is a random function $B(x, x^*)$ of x and x^* that does not depend on z . We write $E[B(x, x^*)]$ as $b(x, x^*)$.

This assumption states that, at the individual level, the magnitude of the direct effect of x compared with x^* on the outcome Y is the same on an additive scale for all z . A detailed mechanistic discussion of this assumption is given in R&G (1992). As noted by Pearl (2001), this assumption is satisfied in

the usual linear SEM model and has been used to identify direct and indirect effects in the structural equation literature. Indeed, in the linear SEM model, $Y(x, z) - Y(x^*, z)$ is usually assumed to be a deterministic function of x and x^* . The following theorem shows that direct and indirect effects are identified by a FRCISTG model under the no-interaction assumption. As in Pearl (2001), we generalize our definitions to non-dichotomous treatments by defining the effects of x compared with x^* as follows:

$$\begin{aligned} \text{PDE}(x, x^*) &= \text{E}[Y\{x, Z(x^*)\}] - \text{E}[Y(x^*)], \\ \text{TIE}(x, x^*) &= \text{E}[Y(x)] - \text{E}[Y\{x, Z(x^*)\}], \\ \text{PIE}(x, x^*) &= \text{E}[Y(x^*, Z(x))] - \text{E}[Y(x^*)], \\ \text{TDE}(x, x^*) &= \text{E}[Y(x)] - \text{E}[Y(x^*, Z(x))]. \end{aligned}$$

These new definitions reduce to the old on choosing $x = 1$ and $x^* = 0$.

Theorem 2.1 *Under the no-interaction assumption,*

$$\begin{aligned} \text{PDE}(x, x^*) &= \text{TDE}(x, x^*) = b(x, x^*), \\ \text{PIE}(x, x^*) &= \text{TIE}(x, x^*), \text{ and} \\ \text{TDE}(x, x^*) + \text{TIE}(x, x^*) &= \text{E}[Y(x)] - \text{E}[Y(x^*)]. \end{aligned}$$

Given data on V , all these quantities are identified in a FRCISTG causal model.

Proof It follows immediately from the no-interaction assumption that

$$\text{PDE}(x, x^*) = \text{TDE}(x, x^*) = \text{E}[B(x, x^*)] = b(x, x^*).$$

The equality $\text{PDE}(x, x^*) = \text{TDE}(x, x^*)$ immediately implies both $\text{PIE}(x, x^*) = \text{TIE}(x, x^*)$ and $\text{TDE}(x, x^*) + \text{TIE}(x, x^*) = \text{E}[Y(x)] - \text{E}[Y(x^*)]$. Finally, $\text{PIE}(x, x^*) = \text{E}[Y(x, z)] - \text{E}[Y(x^*, z)]$, $\text{E}[Y(x)]$, and $\text{E}[Y(x^*)]$ are identified in an FRCISTG model. \square

It seems biologically rather unlikely that the no-interaction assumption will hold when Z affects Y . The no-interaction assumption can be tested in an FRCISTG model since it implies the testable restriction that $\text{E}[Y(x, z) - Y(x^*, z)]$ does not depend on z . More realistic assumptions with weaker consequences are considered in the following theorem. We say that X and Z never interact negatively if $x > x^*$ implies $Y\{x, z\} - Y\{x^*, z\}$ is non-decreasing in z . We say that X is non-preventive for Z if $Z(x) \geq Z(x^*)$ when $x > x^*$.

Theorem 2.2 *If X and Z never interact negatively and X is non-preventive for Z , then for $x > x^*$ $\text{TDE}(x, x^*) \geq \text{PDE}(x, x^*)$, $\text{TIE}(x, x^*) \geq \text{PIE}(x, x^*)$, $\text{PIE}(x, x^*) + \text{PDE}(x, x^*) \leq \text{E}[Y(x)] - \text{E}[Y(x^*)] \leq \text{TDE}(x, x^*) + \text{TIE}(x, x^*)$.*

Proof If we can show that $\text{TDE}(x, x^*) - \text{PDE}(x, x^*) \geq 0$, then the remainder of the theorem follows at once from the basic definitions of the quantities

I_2 . For expositional simplicity, we shall assume that on any elaborated graph there is only one path from X to Z as no qualitatively new issues arise when there are multiple paths. Now if DAG 3 is a NPSE model, then so is DAG 1 since the variables (I_1, I_2) being marginalized over are not a common cause (i.e. parent) of any two variables on DAG 1. It is for this reason that when estimating the effect of setting any variable on DAG 1, we neither require data on (I_1, I_2) nor need to include them on DAG 1. Now from DAG 3 we observe that a drug A will succeed in blocking all of X 's effect on Z by blocking the effect of X on I_1 , the effect of I_1 on I_2 , or the effect of I_2 on Z . However, the counterfactual mean of Y were all continuing smokers given drug A can differ in each case and will, as shown in the following paragraph, in general, equal $E[Y \{x = 1, Z(x = 0)\}]$ only if drug A blocks the effect of X on I_1 and I_1 is the unique child of X on the 'maximally elaborated path' from X to Z . We say that a path $X = I_0 \rightarrow I_1 \rightarrow \dots \rightarrow I_{J-1} \rightarrow I_J = Z$, denoted by P , is 'maximally elaborated' if all variables on the causal chain from X to Z are included on P . For the sake of the following argument, assume such a maximally elaborated path exists.

Remark Mathematically we define P to be maximally elaborated if for all $j, 0 \leq j \leq J$, and all variables I^* such that neither (I_j, I^*) nor (I^*, I_{j+1}) have degenerate joint distributions, the DAG that replaces P by the path $X = I_0 \rightarrow I_1 \rightarrow \dots \rightarrow I_j \rightarrow I^* \rightarrow I_{j+1} \dots \rightarrow I_{J-1} \rightarrow I_J = Z$ is not a causal DAG.

Suppose, without loss of generality, that DAG 3 contains the maximally elaborated path from X to Z and drug A blocked the effect of I_1 on I_2 by, as one possibility, forcing I_2 to be $I_2(i_1 = 0)$, say, regardless of the value of I_1 . Then among smokers given A , Z will equal the counterfactual $Z(i_1 = 0)$ and thus the mean of Y is $E[Y(x = 1, i_1 = 0)] = \int E[Y(x = 1, z)] dF_{Z(i_1=0)}(z)$ while, by (2.3), $E[Y \{x = 1, Z(x = 0)\}] = \int E[Y(x = 1, z)] dF_{Z(x=0)}(z)$. These quantities differ because the drug blocks not only the effect of X on Z but also the effect of I_1 on Z and many subjects have non-zero values of I_1 due to non-smoking-related causes that are the source of the NPSE model error term U_{I_1} associated with I_1 . Indeed, if these integrals did not differ, we would have succeeded in identifying $E[Y \{x = 1, Z(x = 0)\}]$ by experimental manipulation of X to 1 and I_1 to 0 on DAG 3, contradicting the fact that $E[Y \{x = 1, Z(x = 0)\}]$ is a non-manipulable parameter. However, if drug A blocked the effect of X on I_1 by making I_1 equal to $I_1(x = 0)$ even when we set X to 1, then the mean of Y among smokers given intervention A is indeed $\int E[Y(x = 1, z)] dF_{Z(x=0)}(z)$. In this case, intervention with A would not correspond to setting or manipulating a variable on causal DAGs 1 or 3. Because it is unlikely that drug A acts on the first link of the maximally elaborated path, we conclude that $E[Y \{x = 1, Z(x = 0)\}]$ and thus the TIE, although possibly of mechanistic interest, will rarely be of direct public health interest, except as an approximation.

What if, following a suggestion of Pearl, we wished to predict the effect on Y of a chemically modified cigarette that totally blocks smoking's ability to elevate

cholesterol. Even in this case, the effect may not be $E[Y \{x = 1, Z(x = 0)\}]$, because the modified cigarette need not necessarily block the first link on the maximally elaborated path from X to Z . However, if we could assume that a modified cigarette is more likely than our drug A to act at an early link on the causal pathway from X to Z , then $E[Y \{x = 1, Z(x = 0)\}]$ would probably better approximate the effect of a modified cigarette on Y than the effect of drug A .

It is not possible to reach agreement on a hypothetical intervention (closest possible world) under which $Y \{x = 1, Z(x = 0)\}$ could be observed, even if we allow for interventions other than the setting of variables, since a maximally elaborated path from X to Z may not exist and is at best ill-defined (in the sense that it is unclear what criteria are to be used in judging a path to be maximally elaborated).

Additional references in discussion

- Arjas, E. and Eerola, M. (1993). On predictive causality in longitudinal studies. *Journal of Statistical Planning and Inference*, **34**, 361–86.
- Eerola, M. (1994). *Probabilistic Causality in Longitudinal Studies*, Lecture Notes in Statistics, No. 92. Springer-Verlag, Berlin.
- Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In *Uncertainty and Artificial Intelligence 11* (eds T. Besnard and S. Hanks), pp. 185–95. Morgan Kaufmann, San Francisco.
- Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, **3**, 405–30.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles, Section 9. Translated in *Statistical Science*, **5**, 465–80, 1990.
- Parner, J. and Arjas, E. (2000). Causal reasoning from longitudinal data. Unpublished manuscript.
- Pearl, J. (2001). Direct and indirect effects. Technical Report R-273, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles.
- Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty and Artificial Intelligence 11* (eds T. Besnard and S. Hanks), pp. 185–95. Morgan Kaufmann, San Francisco.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–512.
- Robins, J. M. (1987a). Errata to ‘A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy

- worker survivor effect.' *Computers and Mathematics with Applications*, **14**, 917-21.
- Robins, J. M. (1987b). Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect.' *Computers and Mathematics with Applications*, **14**, 923-45.
- Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics, No. 120 (ed. M. Berkane), pp. 69-117. Springer-Verlag, New York.
- Robins, J. M. (1998). Structural nested failure time models. In *Encyclopedia of Biostatistics* (eds P. Armitage and T. Colton), pp. 4372-89. Wiley, Chichester.
- Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143-55.
- Robins, J. M. and Greenland, S. (2000). Comment on 'Causal inference without counterfactuals' by A. P. Dawid. *Journal of the American Statistical Society*, **95**, 477-82.
- Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, **6**, 34-58.
- Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.