

In: Highly Structured Stochastic Systems,  
Eds by P. Green, N. Hjort and S. Richardson,  
Oxford University Press, 2002 (to appear).

James M. Robins  
Semantics of Causal DAG Models and The  
Identification of Direct and Indirect Effects

*Harvard School of Public Health*

Directed acyclic graphs (DAGs) are commonly used to represent causal models. The paper by Dawid in this volume posits a causal model that is closely related to the model of Spirtes et al. (1993) and the model of Pearl (1993). In this discussion I will compare and contrast the semantics of DAGs representing the Spirtes et al. model to that of DAGs representing the non-parametric structural equation (NPSEM) model of Pearl (1995) 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. To give the flavor of the issues involved I will review in detail one such 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 various causal models to be compared. In Section 2, I collect mathematical results on the identification of direct and indirect effects. In Section 3, I discuss substantive implications of these results. In Section 4, I discuss Phil Dawid's paper.

## 1 Causal models and their DAG representation

We are given a DAG  $G$  with 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 now consider three causal models that may be represented by the DAG  $G$ . Let  $X$  denote any subset of  $V$  and let  $x$  be a realization of  $X$ . Two of the causal models assume the existence of counterfactuals  $V_m(x)$  where  $V_m(x)$  is the random variable encoding the value the variable  $V_m$  would have if, possibly contrary to fact,  $X$  were set to  $x$  and  $V_m(x)$  is assumed to be well defined in the sense that there is reasonable agreement as to the hypothetical intervention (i.e., closest possible world) which sets  $X$  to  $x$  (Robins and Greenland, 2000). We shall use the following notational conventions. For any variable  $Z$ , let  $\mathcal{Z}$  be the support (i.e., the set of possible realizations) of  $Z$ . For any  $z_0, \dots, z_m$ , define  $\bar{z}_m = (z_0, \dots, z_m)$ . By convention  $\bar{z}_{-1} \equiv z_{-1} \equiv 0$ .

A finest FR CISTG model assumes (i) all one-step ahead counterfactuals  $V_m(\bar{v}_{m-1})$  exist, (ii)  $V_m(\bar{v}_{m-1}) \equiv V_m(pa_{m-1})$  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  $V_3(v_1) = V_3\{v_1, V_2(v_1)\}$ ,  
(iv)

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \prod V_m | \bar{V}_{m-1} = \bar{v}_{m-1} \quad (1)$$

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

A NPSEM 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.

The relationship between NPSEMs and finest FR CISTGs is given in the following.

**Lemma 1:** A NPSEM can be equivalently characterized by (i) - (iv) under the definition of a FR CISTG except with Eq. (1) replaced by

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \prod V_m | \bar{V}_{m-1} = \bar{v}_{m-1}^* . \quad (2)$$

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

**Remark:** In my 1995 Biometrika comment on Pearl (1995), I incorrectly claimed in my Lemma 1 that a NPSEM and a finest FR CISTG were equivalent. Butch Tsiatis pointed out to me that I had failed to note that an NPSEM satisfied the stronger assumption of Eq. (2). The proof of the (corrected) Lemma 1 proceeds as in my Biometrika comment on Pearl (1995) where the following Lemma was also proved.

**Lemma 2:** If a DAG  $G$  represents a FRCISTG, then the density  $f_V(V)$  of the observables  $V$  satisfies the Markov factorization

$$f_V(v) = \prod_{j=1}^M f(v_j | pa_j) . \quad (3)$$

**Intervention distributions on FR CISTGs:** Suppose we are given a set of variables  $X = \{X_1, \dots, X_k\} \subset V$  and a DAG  $G(X)$  that agrees with DAG  $G$  except the parents  $PA_{G(X),m}$  of  $X_m \in X$  may differ from the parents of  $X_m$  on  $G$ . A non-random  $G(X)$  - specific treatment regime  $g_{G(X)}$  is a collection of functions  $g_{G(X)} = \{g_{G(X),0}, \dots, g_{G(X),k}; g_{G(X),m} : PA_{G(X),m} \rightarrow \mathcal{X}_m\}$  that gives the value  $g_{G(X),m}(pa_{G(X),m})$  that we will set  $X_m$  to when  $PA_{G(X),m} = pa_{G(X),m}$ . When for each  $m$ ,  $X_m$  has no parents on  $G(X)$ , so that  $g_{G(X),m}(pa_{G(X),m})$  is a constant, say  $x_m^*$ , we say regime  $g_{G(X)}$  is non-dynamic and write  $g_{G(X)} = x^* = \{x_1^*, \dots, x_k^*\}$ . Otherwise,  $g_{G(X)}$  is dynamic. The counterfactual random variable  $V_j(g_{G(X)})$  associated with regime  $g_{G(X)}$  is recursively defined to be the one step ahead counterfactual  $V_j(\bar{v}_{j-1})$  evaluated at  $\bar{v}_{j-1} = \bar{V}_{j-1}(g_{G(X)})$  when  $V_j \in V \setminus X$  and at  $V_j(\bar{v}_{j-1}) = g_{G(X),m}(pa_{G(X),m})$  with  $pa_{G(X),m} = PA_{G(X),m}(g_{G(X)})$  when  $V_j = X_m \in X$ .

**Lemma 3 (Robins, 1986):** If DAG  $G$  represents a FRCISTG, then for any set of variables  $X \subset V$ , any DAG  $G(X)$ , and any treatment regime  $g_{G(X)}$ , the (intervention) density  $f_{V(g_{G(X)})}(v)$  of the counterfactual  $V(g_{G(X)})$  is a functional of  $f_V(v)$  and thus is non-parametrically identified. This functional, which I have referred to as the  $g$ -computation algorithm functional or density (hereafter  $g$ -functional or density), is the density  $f_{g_{G(X)}}(v)$  obtained by modifying the product on the right-hand side of (3) as follows: for  $v_j = x_m$  with  $X_m \in X$  remove the term  $f(v_j | pa_j)$  from the product and set  $v_j = x_m$  to  $g_{G(X),m}(pa_{G(X),m})$  elsewhere.

The third causal model that we shall consider simply assumes that the joint distribution of  $V$  factors as in (3) and that the joint density of  $V$  under the intervention  $g_{G(X)}$  is  $f_{g_{G(X)}}(v)$ . It is a simplified version of the causal DAG models of Sprites et al. (1993) and Pearl (1993). Although this model assumes that density of  $V$  under the intervention  $g_{G(X)}$  is well defined, the model makes no reference to counterfactual variables and is agnostic as to their existence. We henceforth refer to this model as the agnostic causal model.

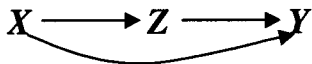
**Remark:** All three causal models assume that the intervention distribution under regime  $g_{G(X)}$  is the  $g$ -functional  $f_{g_{G(X)}}(v)$  of  $f_V(v)$ . However, it generally happens, as in Phil Dawid's paper, that we are only interested in the marginal intervention distribution  $f_{g_{G(X)}}(y) = \int \cdots \int f_{g_{G(X)}}(v) d\mu(y^c)$  of a subset  $Y$  of the variables in  $V = (Y, Y^c)$ , say, and further that data are obtained only on a subset  $V^*$  of the variables in  $V$  with  $Y \subset V^*$ . In this case one wishes to know whether the intervention distribution  $f_{g_{G(X)}}(y)$  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  $f_{g_{G(X)}}(y)$  as discussed by Dawid. 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 (1995) 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. For example, the finest FRCISTG model is a manipulative model since the non-parametrically identified causal parameters are all functions of the  $f_{V(g_{G(X)})}(v)$  which can be tested, in principle, by manipulation. 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 regime  $g_{G(X)}$ , allowing us to estimate the intervention distribution. If, for some  $g_{G(X)}$ , the g-functional  $f_{g_{G(X)}}(v)$  differs from the intervention distribution, 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., 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. We shall now show that, in contrast to the finest FRCISTG and agnostic models, the NPSEM model is a non-manipulative model. We will do so in the context of the estimation of direct and indirect effects of a treatment.



DAG 1

Consider the complete DAG 1. 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 NPSEM or FRCISTG model,

$$\begin{aligned} PDE &= E[Y\{x=1, Z(x=0)\}] - E[Y(x=0)] = \\ &E[Y\{x=1, Z(x=0)\}] - E[Y(x=0, Z(x=0))] \end{aligned} \quad (3)$$

since  $E[Y(x=0)] = E[Y(x=0, Z(x=0))]$ . Here  $Y(x, z)$  is the counterfactual  $Y$  with  $(X, Z)$  set to  $(x, z)$  and  $Z(x)$  is the counterfactual  $Z$  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,

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

since  $E[Y(x=1)] - E[Y(x=0)]$  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 NPSEM or FRCISTG,

$$PIE = E[Y(x=0, Z(x=1))] - E[Y(x=0)]. \quad (4)$$

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 PIE. Thus,

$$TDE = E[Y(x=1)] - E[Y(x=0, Z(x=1))].$$

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  $E[Y\{x=1, Z(x=0)\}]$  and  $E[Y(x=0, Z(x=1))]$  are not assumed to exist.

The direct effect of  $X$  when  $Z$  is set to  $z$  (i.e.,  $E[Y(1, z)] - E[Y(0, z)]$ ) is identified from  $f_V(v)$  by the g-formula under all three models. But, in general, the contrast  $E[Y(1, z)] - E[Y(0, z)]$  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  $E[Y(x)]$  is identified under any of the three causal models, determining whether  $TIE$ ,  $PIE$ ,  $TDE$  and  $PDE$  are identified is equivalent to determining whether  $E[Y(x, Z(x^*))]$ ,  $x \neq x^*$  is identified. Pearl (2001) showed that, if there exists a set  $W \subset V$  of non-descendants of  $X$  and  $Z$ , such that

$$Y(x, z) \perp\!\!\!\perp Z(x^*) \mid W \text{ for all } z \quad (5)$$

then  $E[Y(x, Z(x^*))]$  equals

$$\iint E[Y(x, z) \mid W = w] dF_{Z(x^*)}(z \mid W = w) dF_W(w). \quad (6)$$

Now Eq. (6) is non-parametrically identified from  $f_V(v)$  under all three causal models since  $E[Y(x, z) \mid W = w]$  and  $f_{Z(x^*)}(z \mid W = w)$  are identified by the g-formula. However, no FRCISTG model implies (5). For an NPSEM, (5) will hold if and essentially only if there is no descendant of  $X$  that is ancestor of both  $Z$  and  $Y$ . Thus, under a NPSEM,  $PIE$  and  $PDE$  will be non-parametrically identified based on DAG 1 but would not be identified based on DAG 2. Note that, regardless of whether or not it is equal to  $E[Y(x, Z(x^*))]$ , the non-parametrically identified parameter (6) always has the interpretation of the mean of  $Y$  when  $X$  is set to  $x$  and  $Z$  is randomly assigned to subjects with probability  $f_{Z(x^*)}(z \mid W = w)$ .



DAG 2

**A Non-manipulative Model:** We now turn to the question of whether  $E[Y(x, Z(x^*))]$  can be identified by manipulation (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 will usually 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 can not 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 [i.e.,  $Z(x) = Z(x^*)$ ]. We will thus refer to  $E[Y(x, Z(x^*))]$  and the pure and total, direct and indirect effect contrasts as non-manipulative parameters. (In those rare cases where such “return to the pre-intervention state” is possible, R&G (1992) note that we estimate  $E[Y(x, Z(x^*))]$  from data obtained in randomized crossover trials with a sufficiently long washout-interval to insure no carry-over effects between treatment periods.)

From the above we conclude that NPSEM causal model in contrast to the agnostic and FRCISTG causal models is a non-manipulative model. In section 2 we implicitly argued that manipulative models are preferable because the predictions of non-manipulative models are not, even in principle, testable by experiment. Here are some counterarguments defending the use of non-manipulative models.

First, the assumptions required to identify  $E[Y(x, Z(x^*))]$  without data from a cross-over trial 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 identify (i)  $E[Y(x)] - E[Y(x^*)]$  from non-randomized data and (ii)  $E[Y(x, Z(x^*))]$  from non-cross over trial data if we are willing to assume in case (i) that  $Y(j) \perp\!\!\!\perp X \mid W, j = 0, 1$  for some *nondescendant*  $W$  of  $X$  and  $Y$  and in case (ii) that Eq. (5) holds. 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 lack of exchangeability. (Pearl (2000) effectively assumes that one can always take a simple random sample of a population and use this sample as subjects of an experiment and have the remainder of the population serve as subjects in an observational study, thereby assuring exchangeability. However, in practice this would rarely be the case.) Fourth, suppose one is working with an FRCISTG with  $V = (X, Z, Y)$  so that

$$Y(x, z) \perp\!\!\!\perp Z(x) \mid X = x \text{ for all } z, x. \quad (7)$$

Letting  $W = X$ , it is hard to construct realistic (as opposed to mathematical) scenarios in which one would accept (7) but not (5) as true: it is unlikely one would accept as true either (7) or (5) unless one believed that  $Z = Z(X)$  was effectively randomly assigned by nature within levels of  $X$ , in which case both (5) and (7) would be true.

**An Alternative Identifying Assumption:** Even if one were to buy the above 4 arguments for using an NPSEM 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 (5) that there is no variable, say  $U$ , which 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, 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 to  $x^*$  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. We first generalize our definitions to non-dichotomous treatments by defining effects of  $x$  compared to  $x^*$  as follows:  $PDE(x, x^*) = E[Y\{x, Z(x^*)\}] - E[Y(x^*)]$ ,  $TIE(x, x^*) = E[Y(x)] -$

$E[Y\{x, Z(x^*)\}]$ ,  $PIE(x, x^*) = E[Y(x^*, Z(x))] - E[Y(x^*)]$ ,  $TDE(x, x^*) = E[Y(x)] - E[Y(x^*, Z(x))]$ . These new definitions reduce to the old on choosing  $x = 1$  and  $x^* = 0$ .

**Theorem:** Under the no-interaction assumption,  $PDE(x, x^*) = TDE(x, x^*) = b(x, x^*)$ ,  $PIE(x, x^*) = TIE(x, x^*)$ , and  $TDE(x, x^*) + TIE(x, x^*) = E[Y(x)] - E[Y(x^*)]$ . Further all these quantities are identified in a FRCISTG causal model.

**Proof:** It follows immediately from the no-interaction assumption that  $PDE(x, x^*) = TDE(x, x^*) = E[B(x, x^*)] = b(x, x^*)$ . The equality  $PDE(x, x^*) = TDE(x, x^*)$  immediately implies both  $PIE(x, x^*) = TIE(x, x^*)$  and  $TDE(x, x^*) + TIE(x, x^*) = E[Y(x)] - E[Y(x^*)]$ . Finally,  $b(x, x^*) = E[Y(x, z)] - E[Y(x^*, z)]$ ,  $E[Y(x)]$ , and  $E[Y(x^*)]$  are identified in an FRCISTG model.

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 restrictions that  $E[Y(x, z) - Y(x^*, z)]$  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:** If  $X$  and  $Z$  never interact negatively and  $X$  is non-preventive for  $Z$  then for  $x > x^*$   $TDE(x, x^*) \geq PDE(x, x^*)$ ,  $TIE(x, x^*) \geq PIE(x, x^*)$ ,  $PIE(x, x^*) + PDE(x, x^*) \leq E[Y(x)] - E[Y(x^*)] \leq TDE(x, x^*) + TIE(x, x^*)$ .

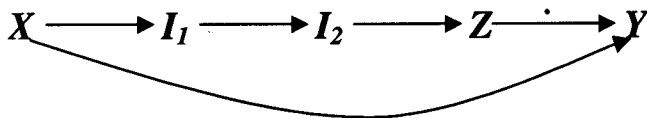
**Proof:** If we can show that  $TDE(x, x^*) - PDE(x, x^*) \geq 0$ , the remainder of the theorem follows at once from the basic definitions of the quantities involved. Now  $TDE(x, x^*) - PDE(x, x^*) = E\{Y(x, Z(x)) - Y(x^*, Z(x))\} - E\{Y(x, Z(x^*)) - Y(x^*, Z(x^*))\} \geq 0$  because  $Y(x, Z(x)) - Y(x^*, Z(x)) \geq Y(x, Z(x^*)) - Y(x^*, Z(x^*))$  under the suppositions of the theorem.

### 3 Substantive Considerations

In this section we investigate through a particular example whether and when scientific interest might lie in estimation of pure and total direct and indirect effects, regardless of whether they are identifiable. R&G (1992) consider a setting in which  $X$  is smoking,  $Z$  is hypercholesterolemia, and  $Y$  is cardiovascular disease, and data are available from a randomized trial of smoking cessation. For simplicity we take all variables to be dichotomous (0,1) variables and assume any noncompliance to be completely at random. We let  $Y = 1$ ,  $Z = 1$ , and  $X = 1$  denote the presence of cardiovascular disease, hypercholesterolemia, and continued smoking. We assume the no-interaction assumption is false because in some subjects hypercholesterolemia produced coronary artery stenosis (narrowing), the narrowed artery was blocked by a blood clot caused by smoking-induced platelet aggregation, and the blocked artery resulted in a heart attack. Thus the smoking effect  $E[Y(x=1)] - E[Y(x=1, z=0)]$  that could be eliminated by controlling all hypercholesterolemia (i.e., setting  $z$  to 0) differs from the total indirect effect (TIE) of smoking  $E[Y(x=1)] - E[Y\{x=1, Z(x=0)\}]$ . We

will assume that smoking and cholesterol never interact negatively and smoking is non-preventive for hypercholesterolemia. We are interested in possible adjuvant treatments for smokers who cannot or will not stop smoking. R&G state that  $E[Y(x=1)] - E[Y(x=1, z=0)]$  would be the parameter of public health policy interest whenever (1) there exists an adjuvant therapy that effectively controls hypercholesterolemia (i.e., a cholesterol lowering drug) but (2) there is no intervention that specifically blocks smoking's effect on cholesterol. R&G go on to say that if a cholesterol lowering drug was unavailable but there was a drug that specifically blocked smoking's ability to elevate cholesterol, then it would be the TIE that would be of public health interest.

We now argue that R&G's final statement is correct only as an approximation. We will assume that a NPSEM represented by DAG 1 is the true state of nature so that Eq. (5) holds and thus pure and total, direct and indirect effects are identified via (6). Further we assume there is a drug  $A$  that completely blocks the effect of smoking  $X$  on cholesterol  $Z$  but does not affect the direct effect of smoking on  $Y$ . If R&G's final statement were precisely correct then, were all continuing smokers in the trial given the drug  $A$ , the mean of  $Y$  would be  $E[Y\{x=1, Z(x=0)\}]$  drug. We now show this need not be the case.



DAG 3

Essentially, any causal pathway  $X \rightarrow Z$  can be further elaborated by adding to the DAG variables that mediate the effect of  $X$  on  $Z$ . For example on DAG 3, the effect  $X \rightarrow Z$  is shown to be through the intermediates  $I_1$  and  $I_2$ . 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 NPSEM 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 do not require data on and thus a DAG that includes  $(I_1, I_2)$ . 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  $P$ . Formally  $P$  is "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. For the time being we will assume such a maximally elaborated path exists.

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 setting  $I_1$  to 0, say. Then the mean of  $Y$  among smokers given  $A$  is  $\int E[Y(x=1, z)] dF_{Z(i_1=0)}(z)$  while, by (6),  $E[Y\{x=1, Z(x=0)\}] = \int E[Y(x=1, z)] dF_{Z(x=0)}(z)$ . These quantities differ because the drug blocks not only the effect of  $X$  on  $Z$  but also the effect of  $I_1$  on  $Z$ . Indeed if these quantities did not differ we would have succeeded in identifying  $E[Y\{x=1, Z(x=0)\}]$  by experimental manipulation of  $I_1$  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  $\int E[Y(x=1, z)] dF_{Z(i_1=0)}(z)$ . This does not contradict our previous results since intervention with  $A$  would not correspond to setting or manipulating a variable on causal DAGs 1 or 3.

**Remark:** Note that the  $Z$ -residual  $U_Z^1$  on DAG 1 is the vector  $(U_{I_1}^3, U_{I_2}^3, U_Z^3)$  where, for example,  $U_Z^3$  is the DAG 3  $Z$ -residual and the DAG 1 and DAG 3  $Z$ -functions are related by  $f_Z^1(X, U_Z^1) = f_Z^3(f_{I_2}(f_{I_1}(X, U_{I_1}^3), U_{I_2}^3), U_Z^3)$ . Interestingly, from the perspective of DAG 1, if drug  $A$  blocked the effect of  $I_1$  on  $I_2$  by setting  $I_1$  to 0, the drug changes the  $Z$ -residual  $U_Z^1$  on DAG 1 from  $(U_{I_1}^3, U_{I_2}^3, U_Z^3)$  to  $(U_{I_2}^3, U_Z^3)$  and the function  $f_Z^1(x=0, U_Z^1)$  from  $f_Z^3(f_{I_2}(f_{I_1}(x=0, U_{I_1}^3), U_{I_2}^3), U_Z^3)$  to  $f_Z^3(f_{I_2}(i_1=0, U_{I_2}^3), U_Z^3)$ . Thus, in contrast to the usual intervention in which a variable on the causal DAG 1 in question is set to a particular value, the drug  $A$  intervention changes both the residual  $U_Z^1$  and the function  $f_Z^1(x=0, U_Z^1)$  on causal DAG 1.

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), we conclude that  $E[Y\{x=1, Z(x=0)\}]$  and thus TIE, although possibly of mechanistic interest, may never be of direct public health interest, except possibly as an approximation. Indeed, because of the maximally elaborated path being ill-defined, 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.

#### 4 Dawid and Causal Decision Theory

In his paper, 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 (1987) 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. It follows that for Dawid pure and total, direct and indirect effects are undefined because their definition requires the existence of counterfactuals as their definition is in terms of the joint distribution of  $Z(x)$  and  $Z(x^*)$ . In contrast, provided that  $X$  and  $Z$  are potentially manipulatable, the direct effect of  $x$  compared to  $x^*$  when  $Z$  is set to a  $z$  is well-defined under Dawid's model and is identified by the g-formula if data on all variables on his causal DAG were available. However as mentioned previously Dawid's focus was on the case where causal contrasts were not identified because data on some variables on his DAG were unavailable.

## References

- Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In: **Uncertainty and Artificial Intelligence**, T. Besnard and S. Hanks, Eds. Vol. 11, pgs. 185-195. San Francisco, CA: Morgan Kaufmann.
- Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, **3**, 405-430.
- Pearl, J. (1993). Comment: Graphical models, causality and interventions. *Statistical Science*, **8**, 266-269.
- Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82** (4), 669-710.
- Pearl, J. and Robins, J.M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. *Uncertainty in Artificial Intelligence. Proceedings of the 11th Conference*, pp. 444-453.
- Pearl, J. (2000). **Causality**. Cambridge, England: Cambridge University Press.
- Pearl, J. (2001). *Technical Report R-273*. Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles.
- 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**, 1393-1512.
- Robins, J.M. (1987). 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** (9-12), 923-945.
- Robins, J.M. (1987). 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**, 917-921.
- Robins, J.M. (1989). Errata to 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*, **18**, 477.

Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143-155.

Robins, J.M. (1997). Causal inference from complex longitudinal data. In: **Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics (120)**. M. Berkane, Editor. New York: Springer Verlag. pp. 69-117.

Robins, J.M. and Greenland, S. (2000). Comment on Causal inference without counterfactuals by A.P. Dawid. *The Journal of the American Statistical Society - Theory and Methods*, **95**, (450) 477-482.

Spirtes, P., Glymour, C., and Scheines, R. (1993). **Causation, Prediction, and Search**. New York: Springer Verlag.