

Alternative Graphical Causal Models and the
Identification of Direct Effects

James Robins

Thomas Richardson

Harvard School of Public Health

University of Washington

Working Paper no. 100

Center for Statistics and the Social Sciences

University of Washington

29 March, 2010

Abstract

We consider four classes of graphical causal models: the Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) of Robins (1986), the agnostic causal model of Spirtes et al. (1993), the Non-Parametric Structural Equation Model (NPSEM) of Pearl (2000), and the Minimal Counterfactual Model (MCM) which we introduce. The latter is referred to as ‘minimal’ because it imposes the minimal counterfactual independence assumptions required to identify those causal contrasts representing the effect of an ideal intervention on any subset of the variables in the graph. The causal contrasts identified by an MCM are, in general, a strict subset of those identified by a NPSEM associated with the same graph. We analyze various measures of the ‘direct’ causal effect, focussing on the pure direct effect (PDE), also called the ‘natural direct effect’. We show the PDE is a parameter that may be identified in a DAG viewed as a NPSEM, but not as an MCM. In spite of this, Pearl has given a scenario in which the PDE corresponds to the intent-to-treat parameter of a randomized experiment. We resolve this apparent paradox by showing that implicit within Pearl’s account is an extended causal DAG with additional variables in which there is a causal contrast that equals the pure direct effect of Pearl’s original NPSEM. Further, this contrast is identified from observational data on the original variables. Finally we relate our results to the work of Avin et al. (2005) on path-specific causal effects.

This paper will appear in *Causality and psychopathology: finding the determinants of disorders and their cures*, P. Shrout, Editor.

1 Introduction

The subject-specific data from either an observational or experimental study consist of a string of numbers. These numbers represent a series of empirical measurements. Calculations are performed on these strings and causal inferences are drawn. For example, an investigator might conclude that the analysis provides strong evidence for “both an indirect effect of cigarette smoking on coronary artery disease through its effect on blood pressure and a direct effect not mediated by blood pressure.” The nature of the relationship between the sentence expressing these causal conclusions and the statistical computer calculations performed on the strings of numbers has been obscure. Since the computer algorithms are well-defined mathematical objects, it is crucial to provide formal causal models for the English sentences expressing the investigator’s causal inferences. In this paper we restrict ourselves to causal models that can be represented by a directed acyclic graph.

There are two common approaches to the construction of causal models. The first approach posits unobserved fixed ‘potential’ or ‘counterfactual’ outcomes for each unit under different possible joint treatments or exposures. The second approach posits relationships between the population distribution of outcomes under experimental interventions (with full compliance) to the set of (conditional) distributions that would be observed under passive observation (i.e., from observational data). We will refer to the former as ‘counterfactual’ causal models and the latter as ‘agnostic’ causal models (Spirtes et al., 1993), as the second approach is agnostic as to whether unit-specific counterfactual outcomes exist, be they fixed or stochastic.

The primary difference between the two approaches is ontological: the counterfactual approach assumes that counterfactual variables *exist*, while the agnostic approach does not require this. In fact, the counterfactual theory logically subsumes the agnostic theory in the sense that the counterfactual approach is logically an extension of the former approach. In particular, for a given graph the causal contrasts (i.e. parameters) that are well-defined under the agnostic approach are also well-defined under the counterfactual approach. This set of contrasts corresponds to the set of contrasts between treatment regimes (strategies) which could be implemented in an experiment with sequential treatment assignments (ideal

interventions), wherein the treatment given at stage m is a (possibly random) function of past covariates on the graph. We refer to such contrasts or parameters as ‘manipulable with respect to a given graph’. As discussed further in Section 2.8, the set of manipulable contrasts for a given graph are identified under the associated agnostic causal model from observational data with a positive joint distribution and no hidden (i.e. unmeasured) variables. A parameter is said to be identified if it can be expressed as a known function of the distribution of the observed data. A discrete joint distribution is positive if the probability of a joint event is nonzero whenever the marginal probability of each individual component of the event is nonzero.

Although the agnostic theory is contained within the counterfactual theory, the reverse does not hold. There are causal contrasts that are well-defined within the counterfactual approach that have no direct analog within the agnostic approach. An example that we shall discuss in detail is the pure direct effect (aka, a natural direct effect) introduced in Robins and Greenland (1992). The pure direct effect of a binary treatment X on Y relative to an intermediate variable Z is the effect the treatment X would have had on Y had (contrary to fact) the effect of X on Z been blocked. The pure direct effect is non-manipulable relative to X , Y and Z in the sense that, in the absence of assumptions, the pure direct effect does not correspond to a contrast between treatment regimes of any randomized experiment performed via interventions on X , Y and Z .

In this paper we discuss three counterfactual models, all of which agree in two important respects: First they agree on the set of well-defined causal contrasts; second they make the consistency assumption that the effect of a (possibly joint) treatment on a given subject depends neither on whether the treatment was freely chosen by, versus forced on, the subject nor on the treatments received by other subjects. However the counterfactual models do not agree as to the subset of these contrasts that can be identified from observational data with a positive joint distribution and no hidden variables. Identifiability of causal contrasts in counterfactual models is obtained by assuming that (typically, conditional on prior history), the treatment received at a given time is independent of some set of counterfactual outcomes. Different versions of this independence assumption are possible: the stronger the assumption (i.e., the more counterfactuals assumed independent of treatment), the more causal contrasts

that are identified. For a given graph G , all the counterfactual models we shall consider identify the set of contrasts identified under the agnostic model for G . We refer to this set of contrasts as *the manipulable contrasts relative to G* .

Among the counterfactual models we discuss, the model derived from the non-parametric structural equation model (NPSEM) of Pearl (2000) makes the strongest independence assumption; indeed the assumption is sufficiently strong that the pure direct effect may be identified (Pearl, 2001). In contrast, under the weaker independence assumption of the Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) counterfactual model of Robins (1986) or the Minimal Counterfactual Model (MCM) introduced in this paper, the pure direct effect is not identified. The MCM is the weakest counterfactual model (i.e., contains the largest set of distributions over counterfactuals) that satisfies the consistency assumption and identifies the set of manipulable contrasts based on observational data with a positive joint distribution and no hidden variables. The MCM is equivalent to the FFRCISTG model, when all variables are binary. Otherwise the MCM is obtained by a mild further weakening of the FFRCISTG independence assumption.

The identification of the non-manipulable pure direct effect parameter under a NPSEM appears to violate the slogan “no causation without manipulation”. Indeed, Pearl has recently advocated the alternative slogan “causation before manipulation” in arguing for the ontological primacy of causation relative to manipulation Pearl (2010); an ontological primacy that follows, for instance, from the philosophical position that all dependence between counterfactuals associated with different variables is due to the effects of common causes (that are to be included as variables in the model and on the associated graph G), thus privileging the NPSEM over other counterfactual models. Pearl (2000) privileges the NPSEM over other models but presents different philosophical arguments for his position.

Pearl’s view is an anathema to those with a positivist, Popperian view of causality (e.g., Dawid (2000a)) who argue that a theory that allows identification of non-manipulable parameters (relative to a graph G) is not a scientific theory because some of its predictions (e.g., that the PDE is a particular function of the distribution of the observed data) are not experimentally testable and thus are non-refutable. Indeed, in appendix C, we give an example of a data generating process that satisfies the assumptions of an FFRCISTG model

but not those of an NPSEM such that (i) the NPSEM prediction for the PDE is false but (ii) the predictions made by all four causal models for the manipulable parameters relative to the associated graph G are correct. In this setting, anyone who assumed an NPSEM would falsely believe she was able to consistently estimate the PDE parameter from observational data on the variables on G and no possible experimental intervention on these variables could refute either her belief in the correctness of the NPSEM or her belief in the validity (i.e., consistency) of her estimator of the PDE. In Appendix C, we derive sharp bounds for the PDE under the assumption that FFRCISTG model associated with graph G holds. We find that these bounds may be quite informative, even though the PDE is not (point) identified.

This strict positivist view of causality relies on the belief that there is a sharp separation between the manipulable and non-manipulable causal contrasts (relative to graph G) because every prediction made concerning a manipulable contrast based on observational data can be checked by an experiment involving interventions on variables in G . However, this view ignores the fact that (i) such experiments may be infeasible or unethical; (ii) such empirical experimental tests will typically require an auxiliary assumption of exchangeability between the experimental and observational population and the ability to measure all the variables included in the causal model, neither of which may hold in practice; (iii) such tests are themselves based upon the untestable assumption that experimental interventions are ideal. Thus, many philosopher's of science do not agree with the strict positivist's sharp separation between manipulable and non-manipulable causal contrasts.

However, Pearl does not rely on this argument in responding to the positivist critique of the NPSEM that it can identify a contrast, the PDE, that is is not subject to experimental test. Rather, he has responded by describing a scenario in which the pure direct effect associated with a particular NPSEM is identifiable, scientifically meaningful, of substantive interest, and corresponds precisely to the intent to treat parameter of a certain randomized controlled experiment. Pearl's account appears paradoxical in light of the results described above, since it suggests that the pure direct effect may be identified via intervention. Resolving this apparent contradiction is the primary subject of this paper.

We will show that implicit within Pearl's account is a causal model associated with an expanded graph (G') containing more variables than Pearl's original graph (G). Further-

more, although the pure direct effect of the original NPSEM counterfactual model is not a manipulable parameter relative to G , it is manipulable relative to the expanded graph G' . Consequently the pure direct effect is identified by all four of the causal models (agnostic, MCM, FFRCISTG and NPSEM) associated with G' . The existence of this causal model associated with G' , licensed by Pearl's account, constitutes the "additional assumptions" that make the original NPSEM's pure direct effect contrast equal to a contrast between treatments in a randomized experiment – a randomized experiment whose treatments correspond to variables on the expanded graph G' that are absent from the original graph G .

However, the distribution of the variables of the expanded graph G' is not positive. Furthermore, the available data are restricted to the variables of the original graph G . Hence, it is not at all obvious *prima facie* that the expanded causal model's treatment contrasts will be identified.

Remarkably, we prove that, under all four causal models associated with the larger graph G' , the manipulable causal contrast of the expanded causal model that equals the pure direct effect of Pearl's original NPSEM G is identified from observational data on the original variables. This identification crucially relies on certain deterministic relationships between variables in the expanded model. Our proof thus resolves the apparent contradiction; furthermore, it shows that the ontological primacy of manipulation reflected in the slogan "no causation without manipulation" can be maintained by interpreting the pure direct effect parameter of a given counterfactual causal model as a manipulable causal parameter in an appropriate expanded model.

Having said this, although in Pearl's scenario the intervention associated with the expanded causal model was scientifically plausible, we discuss a modification of Pearl's scenario in which the intervention required to interpret the PDE contrast of the original graph G as manipulable contrast of an expanded graph G' is more controversial. Our discussion reveals the scientific benefits that flow from the exercise of trying to provide an interventionist interpretation for a non-manipulable causal parameter identified under an NPSEM associated with a graph G . Specifically, the exercise often helps one devise explicit, and sometimes even practical, interventions, corresponding to manipulable causal parameters of an expanded graph G' . The exercise also helps one recognize when such interventions are quite a stretch.

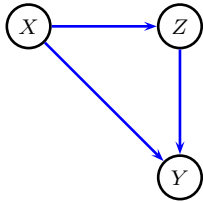


Figure 1: A simple DAG with a treatment X , an intermediate Z and a response Y .

In this paper, our goal is not to advocate for the primacy of manipulation or of causation. Rather our goal is to contribute both to the philosophy and to the mathematics of causation by demonstrating that the apparent conflict between these paradigms is often not a real conflict.

The reduction of an identified non-manipulable causal contrast of a NPSEM to a manipulable causal contrast of an expanded model that is then identified via deterministic relationships under the expanded agnostic model is achieved here for the pure direct effect. A similar reduction for the effect of treatment on the treated (i.e. compliers) in a randomized trial with full compliance in the placebo arm was given by Robins et al. (2007); see also Geneletti and Dawid (2007) and the Appendix herein.

The paper extends and revises previous discussions by Robins and colleagues (Robins, 2003; Robins et al., 2007; Robins and Greenland, 1992) of direct and indirect effects. We restrict consideration to causal models, such as the agnostic, FFRCISTG, MCM, and NPSEM, that can be represented by a directed acyclic graph (DAG). See Robins et al. (2009) for a discussion of alternative points of view due to Hafeman and VanderWeele (2009), Imai et al. (2009) and Petersen et al. (2006) that are not based on DAGs.

The paper is organized as follows: Section 2 introduces the four types of causal model associated with a graph; Section 3 defines direct effects; Section 4 analyzes the conditions required for the ‘pure direct effect’ (PDE) to be identified; Section 5 considers, via examples, the extent to which the PDE may be interpreted in terms of interventions; Section 6 considers path specific effects; finally Section 7 contains the conclusion.

2 Graphical Causal Models

Define a DAG G to be a graph with nodes (vertices) representing the elements of a vector of random variables $V = (V_1, \dots, V_M)$ with directed edges (arrows) and no directed cycles. To

avoid technicalities we assume all variables V_m are discrete. We let $f(v) \equiv f_V(v) \equiv \Pr(V = v)$ all denote the probability density of V , where, for simplicity, we assume $v \in \mathcal{V} \equiv \bar{\mathcal{V}}_M$, $\bar{\mathcal{V}}_m \equiv \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_m$, \mathcal{V}_m denotes the assumed known space of possible values v_m of V_m , and for any z_1, \dots, z_m , we define $\bar{z}_m = (z_1, \dots, z_m)$. By convention for any $\bar{z}_m, \bar{z}_0 \equiv z_0 \equiv 0$. Note $\bar{\mathcal{V}}_m \equiv \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_m$ is the product space of the $\mathcal{V}_j, j \leq m$. We do not necessarily assume that $f(v)$ is strictly positive for all $v \in \mathcal{V}$.

As a simple example, consider a randomized trial of smoking cessation, represented by the DAG G with node set $V = (X, Z, Y)$ in Figure 1. Thus, $M = 3, V_1 = X, V_2 = Z, V_3 = Y$. Here X is the randomization indicator with $X = 0$ denoting smoking cessation and $X = 1$ active smoking, Z is an indicator variable for hypertensive status 1 month post randomization, and Y an indicator variable for having a myocardial infarction by end of follow-up at 3 months. For simplicity assume complete compliance with assigned treatment and assume no subject had an MI prior to 1 month. We refer to the variables V as factual variables as they are variables that could potentially be recorded on the subjects participating in the study. Because in this paper our focus is on identification we assume the study population is sufficiently large that sampling variability can be ignored. Then the density $f(v) = f(x, z, y)$ of the factual variables can be taken to be the proportion of our study population with $X = x, Z = z, Y = y$. Our ultimate goal is to try to determine whether X has a direct effect on Y not through Z .

We use either PA_{V_m} or PA_m to denote the parents of V_m , i.e., the set of nodes from which there is a direct arrow into V_m . For example, in Figure 1, $PA_Y = \{X, Z\}$. A variable V_j is a descendant of V_m if there is a sequence of nodes connected by edges between V_m and V_j such that, following the direction indicated by the arrows, one can reach V_j by starting at V_m , i.e. $V_m \rightarrow \cdots \rightarrow V_j$. Thus, in Figure 1, Z is descendant of X but not of Y .

We suppose that, as in Figure 1, the $V = (V_1, \dots, V_M)$ are numbered so that V_j is not a descendant of V_m for $m > j$.

Let $R = (R_1, \dots, R_K)$ denote any subset of V and let $r = (r_1, \dots, r_K)$ be a value of R . We write $R_j = V_m, \mathcal{R}_j = \mathcal{V}_m$ if the j -th variable in R corresponds to the m -th variable in V . The NPSEM, MCM and FFRCISTG model all assume the existence of the counterfactual random variable $V_m(r)$ encoding the value the variable V_m would have if, possibly contrary

to fact, R were set to r , $r \in \mathcal{R} = \mathcal{R}_1 \otimes \cdots \otimes \mathcal{R}_K$, where $V_m(r)$ 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 R to r (Robins and Greenland, 2000). For example, in Figure 1, $Z(x = 1)$ and $Y(x = 1)$ are a subject’s Z and Y had, possibly contrary to fact, the subject been a smoker. By assumption, if $R_j \in R$ is the m -th variable V_m , then $V_m(r)$ equals the value r_j to which the variable $V_m = R_j$ was set. For example, in Figure 1, the counterfactual $X(x = 1)$ is equal to 1. Note we assume $V_m(r)$ is well-defined even when the factual probability $\text{pr}[R = r]$ is zero. Although we recognize that under certain circumstances such an assumption might be ‘metaphysically suspect’ because the counterfactuals could be ‘radically’ ill-defined, since no one was observed to receive the treatment in question. However, in our opinion, in a number of the examples that we consider in this paper these counterfactuals do not appear to be much less well-defined than those corresponding to treatments that have positive factual probability.

We often write the density $f_{V(r)}(v)$ of $V(r)$ as $f_r^{int}(v)$, with ‘*int*’ short for intervene, to emphasize the fact that $f_{V(r)}(v) = f_r^{int}(v)$ represents the density of V in the counterfactual world where we intervened and set each subject’s R to r . We say that $f_r^{int}(v)$ is the density of V had, contrary to fact, each subject followed the treatment regime r . In contrast, $f(v)$ is the density of the factual variables V .

With this background, we specify our four causal models.

2.1 FFRCISTG Causal Models

Given a DAG G with node set V , a FFRCISTG model associated with G makes four assumptions.

- (i) All one-step ahead counterfactuals $V_m(\bar{v}_{m-1})$ exist for any setting $\bar{v}_{m-1} \in \bar{V}_{m-1}$ of its predecessors.

For example, in Figure 1, a subject’s hypertensive status $Z(x) = V_2(v_1)$ at smoking level x for $x = 0$ and for $x = 1$ exists and a subject’s *MI* status $Y(x, z) = V_3(\bar{v}_2)$ at each joint level of smoking and hypertension exists. Because $V_1 = X$ has no predecessor, $V_1 = X$ only exists as a factual variable.

(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 .

For example, were the edge $X \rightarrow Y$ missing in Figure 1, this assumption would imply that $Y(x, z) = Y(z)$ for every subject and every z . That is, the absence of the edge would imply that smoking X has no effect on Y other than through its effect on Z .

(iii) Both the factual variables V_m and the counterfactuals $V_m(r)$ for any $R \subset V$ are obtained recursively from the $V_m(\bar{v}_{m-1})$. For example, $V_3 = V_3\{V_1, V_2(V_1)\}$ and $V_3(v_1) = V_3\{v_1, V_2(v_1)\}$.

Thus, in Figure 1, with the treatment R being smoking X , a subject's possibly counterfactual MI status $Y(x = 1) = V_3(v_1 = 1)$ had he been forced to smoke is $Y(x = 1, Z(x = 1))$, and thus is completely determined by the one-step ahead counterfactuals $Z(x)$ and $Y(x, z)$. That is, $Y(x = 1)$ is obtained by evaluating the one step ahead counterfactual $Y(x = 1, z)$ at z equal to $Z(x = 1)$. Similarly, a subject's factual X and one-step ahead counterfactuals determine the subject's factual hypertensive status Z and MI status Y as $Z(X)$ and $Y(X, Z(X))$ where $Z(X)$ is the counterfactual $Z(x)$ evaluated at $x = X$ and $Y(X, Z(X))$ is the counterfactual $Y(x, z)$ evaluated at $(x, z) = (X, Z(X))$.

(iv) The following independence holds:

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

for all m and all $\bar{v}_{M-1} \in \bar{V}_{M-1}$

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

Assumption (iv) is equivalent to the statement that for each m , conditional on the factual past $\bar{V}_{m-1} = \bar{v}_{m-1}$, any possible evolution from $m+1$ of one-step ahead counterfactuals (consistent with \bar{V}_{m-1}), i.e., $\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\}$ is independent of the factual variable V_m since by (iii), $V_m \equiv V_m(\bar{V}_{m-1}) = V_m(\bar{v}_{m-1})$ when $\bar{V}_{m-1} = \bar{v}_{m-1}$.

Note that by (iii) above, the counterfactual $V_{m+1}(\bar{v}_m)$ for a given subject, say subject i , depends on the treatment \bar{v}_m received by the subject but does not depend on the treatment

received by any other subject. Further $V_{m+1}(\bar{v}_m)$ takes the same value whether the treatment \bar{v}_m is counter to fact (i.e. $\bar{V}_m \neq \bar{v}_m$) or factual (i.e. $\bar{V}_m = \bar{v}_m$ and thus $V_{m+1}(\bar{v}_m) = V_{m+1}$). That is, the FFRCISTG model satisfies the consistency assumption of the introduction. Indeed, we shall henceforth refer to (iii) as the consistency assumption.

The following example will play a central role in the paper.

Example 1 Consider the FFRCISTG model associated with the graph in Figure 1, then, for all z ,

$$Y(x = 1, z), Z(x = 1) \perp\!\!\!\perp X, \quad Y(x = 0, z), Z(x = 0) \perp\!\!\!\perp X \quad (2)$$

and

$$Y(x = 1, z) \perp\!\!\!\perp Z(x = 1) \mid X = 1, \quad Y(x = 0, z) \perp\!\!\!\perp Z(x = 0) \mid X = 0 \quad (3)$$

are true statements by assumption (iv). However, the model makes no claim as to whether

$$Y(x = 1, z) \perp\!\!\!\perp Z(x = 0) \mid X = 0$$

and

$$Y(x = 1, z) \perp\!\!\!\perp Z(x = 0) \mid X = 1$$

are true, because, for example, the value of x in $Y(x = 1, z)$ differs from the factual value $X = 0$ of X in the conditioning event. We shall see that all the above independence statements are true by assumption under the NPSEM associated with the graph in Figure 1.

2.2 Minimal Counterfactual Models (MCMs)

An MCM differs from a FFRCISTG model only in that (iv) is replaced by:

(iv*) For all m and all $\bar{v}_{M-1} \in \bar{\mathcal{V}}_{M-1}$,

$$\begin{aligned} & f\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}, V_m = v_m\} \\ &= f\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}\}. \end{aligned} \quad (4)$$

Since (iv) can be written as, for all m , all $\bar{v}_{M-1} \in \bar{\mathcal{V}}_{M-1}$, and all v_m^* ,

$$\begin{aligned} & f\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}, V_m = v_m^*\} \\ &= f\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}\}, \end{aligned}$$

condition (iv) for an FFRCISTG implies condition (iv*) for an MCM. However, the reverse does not hold.

An MCM only requires that the last display hold for the unique value v_m of V_m that occurs in the given \bar{v}_{M-1} . Thus, Eq. (4) states that, conditional on the factual past $\bar{V}_{m-1} = \bar{v}_{m-1}$ through $m-1$, any possible evolution from $m+1$ of one-step ahead counterfactuals, $\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\}$, consistent with the past $\bar{V}_m = \bar{v}_m$ through m , is independent of the event $V_m = v_m$. In other words:

$$\begin{aligned} \{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \perp\!\!\!\perp I(V_m(\bar{v}_{m-1}) = v_m) \mid \bar{V}_{m-1} = \bar{v}_{m-1}, \\ \text{for all } m \text{ and all } \bar{v}_{M-1} \in \bar{\mathcal{V}}_{M-1}, \end{aligned} \tag{5}$$

where $I(V_m = v_m)$ is the Bernoulli indicator random variable.

It follows that in the special case where all the V_m are binary, an MCM and an FFRCISTG model are equivalent because, for V_m binary, the random variables V_m and $I(V_m = v_m)$ are the same (possibly up to a recoding).

In Appendix A.2 we describe a data-generating process leading to a counterfactual model that is an MCM associated with the graph $X \rightarrow Y$, but not an FFRCISTG for this graph.

2.3 A representation of MCMs and FFRCISTG models that does not condition on the past

In this section we derive alternative characterizations of the counterfactual independence conditions for the FFRCISTG model and the MCM.

Theorem 1 *Given an FFRCISTG model associated with a graph G :*

- (a) *The set of independencies in Eq. (1) are satisfied if and only if,*

$$\begin{aligned} & \text{for all } \bar{v}_{M-1}, \text{ the random variables} \\ & V_{m+1}(\bar{v}_m), m = 0, \dots, M-1 \text{ are mutually independent.} \end{aligned} \tag{6}$$

(b) *Furthermore the set of independencies (6) is the same for any ordering of the variables compatible with the descendant relationships in G .*

Proof of (a): (\Rightarrow) Given \bar{v}_{M-1} and $m \in \{1, \dots, M-1\}$, we define $\mathfrak{I}_m = \{\mathfrak{I}_{m,m}, \dots, \mathfrak{I}_{M,m}\}$ to be a set of conditional independence statements:

- i) $\mathfrak{I}_{m,m} : \{V_M(\bar{v}_{M-1}), \dots, V_{m+1}(\bar{v}_m)\} \perp\!\!\!\perp V_m(\bar{v}_{m-1})$, and
- ii) for $j = 1$ to $j = M - m$,

$\mathfrak{I}_{m+j,m} :$

$$V_M(\bar{v}_{M-1}), \dots, V_{m+j+1}(\bar{v}_{m+j}) \perp\!\!\!\perp V_{m+j}(\bar{v}_{m+j-1}) \mid \\ V_{m+j-1}(v_{m+j-2}) = v_{m+j-1}, \dots, V_m(\bar{v}_{m-1}) = v_m.$$

First note that the set of independencies in Eq. (1) is precisely \mathfrak{I}_1 . Now, if the collection \mathfrak{I}_m holds (for $m < M - 2$) then \mathfrak{I}_{m+1} holds since (I) the set \mathfrak{I}_{m+1} is precisely the set $\{\mathfrak{I}_{m+1,m}, \dots, \mathfrak{I}_{M,m}\}$ except with $V_m(\bar{v}_{m-1})$ removed from all conditioning events and (II) $\mathfrak{I}_{m,m}$ licenses such removal. Thus, beginning with \mathfrak{I}_1 , we recursively obtain that \mathfrak{I}_m and thus $\mathfrak{I}_{m,m}$ holds for $m = 1, \dots, M - 1$. The latter immediately implies that the variables $V_{m+1}(\bar{v}_m)$, $m = 0, \dots, M - 1$ are mutually independent.

(\Leftarrow) The reverse implication is immediate upon noting that the conditioning event $\bar{V}_{m-1} = \bar{v}_{m-1}$ in Eq. (1) is the event $V_0 = v_0, V_1(v_0) = v_1, \dots, V_{m-1}(\bar{v}_{m-2}) = v_{m-1}$.

Proof of (b): This follows immediately from the assumption that $V_m(\bar{v}_{m-1}) = V_m(pa_m)$. \square

Theorem 2 *Given an MCM associated with a graph G :*

(a) *The set of independencies in Eq. (5) are satisfied if and only if, for all $\bar{v}_{M-1}, m \in 1, \dots, M - 1$,*

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \perp\!\!\!\perp I(V_m(\bar{v}_{m-1}) = v_m). \quad (7)$$

(b) *Furthermore the set of independencies (7) is the same for any ordering of the variables compatible with the descendant relationships in G .*

An immediate corollary is the following.

Corollary 3 *An MCM implies that for all \bar{v}_{M-1} , the random variables $I(V_{m+1}(\bar{v}_m) = v_{m+1})$, $m = 0, \dots, M - 1$ are mutually independent.*

Proof of Theorem 2(a): (\Rightarrow) Given \bar{v}_{M-1} , the proof exactly follows that of the previous theorem when we redefine:

i) $\mathfrak{I}_{m,m} : \{V_M(\bar{v}_{M-1}), \dots, V_{m+1}(\bar{v}_m)\} \perp\!\!\!\perp I(V_m(\bar{v}_{m-1}) = v_m)$, and

ii) for $j = 1$ to $j = M - m$,

$\mathfrak{I}_{m+j,m} :$

$$V_M(\bar{v}_{M-1}), \dots, V_{m+j+1}(\bar{v}_{m+j}) \perp\!\!\!\perp I(V_{m+j}(\bar{v}_{m+j-1}) = v_{m+j}) \mid \\ V_{m+j-1}(v_{m+j-2}) = v_{m+j-1}, \dots, V_m(\bar{v}_{m-1}) = v_m.$$

The reverse implication and (b) follows as in the proof of the previous theorem. \square

2.4 NPSE Causal Models

Given a DAG G with node set V , a NPSEM associated with G assumes that there exist mutually independent random variables ϵ_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 $f_m(pa_m, \epsilon_m)$ and both the factual 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 in (iii) in §2.1.

Under an NPSEM the FFRCISTG condition (1) and the MCM condition (5) both hold. However, an FFRCISTG or MCM associated with G will not, in general, be an NPSEM for G . Indeed a NPSEM implies

$$\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\} \perp\!\!\!\perp V_m(\bar{v}_{m-1}^{**}) \mid \bar{V}_{m-1} = \bar{v}_{m-1}^*, \quad (8)$$

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

That is, conditional on the factual past $\bar{V}_{m-1} = \bar{v}_{m-1}^*$, the counterfactual $V_m(\bar{v}_{m-1}^{**})$ is statistically independent of all future one-step ahead counterfactuals. This implies that all four statements in Example 1 are true under a NPSEM ; see also Pearl (2000, §3.6.3).

Hence, in a MCM or FFRCISTG model, in contrast to a NPSEM, the defining independencies are those for which the value of \bar{v}_{m-1} in (a) the conditioning event, (b) in the counterfactual V_m at m and (c) in the set of future one step ahead counterfactuals $\{V_{m+1}(\bar{v}_m), \dots, V_M(\bar{v}_{M-1})\}$ are equal. Thus a FFRCISTG 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 $\bar{v}_{m-1}^{**} = \bar{v}_{m-1}^* = \bar{v}_{m-1}$. As mentioned above, the MCM further weakens the independence by replacing V_m with $I(V_m = v_m)$.

In Appendix B we describe a data-generating process leading to a counterfactual model that is an MCM / FFRCISTG model associated with Figure 1, but not an NPSEM for this Figure.

Understanding the implications of these additional counterfactual independences assumed by an NPSEM compared to an MCM or FFRCISTG model is one of the central themes of this paper.

2.5 The g-functional

Before defining our third causal model, the agnostic causal model, we need to define the g-functional density. The next Lemma shows that the assumptions of an MCM, and thus *a fortiori* those of the NPSEMs and FFRCISTG models, restrict the joint distribution of the factual variables (when there are missing edges in the DAG).

Lemma 4 *In an MCM associated with DAG G , for all v such that $f(v) > 0$, the density $f(v) \equiv pr(V = v)$ of the factials V satisfies the Markov factorization*

$$f(v) = \prod_{j=1}^M f(v_j | pa_j). \quad (9)$$

Robins (1986) proved Lemma 4 for a FFRCISTG model; the proof applies equally to a MCM. Eq. (9) is equivalent to the statement that each variable V_m is conditionally independent of its non-descendants given its parents (Pearl, 1988).

Example 2 In Figure 1, $f(x, z, y) = f(y | x, z)f(z | x)f(x)$. If the arrow from X to Y were missing, we would then have $f(x, z, y) = f(y | z)f(z | x)f(x)$ since Z would be the only parent of Y .

Definition 5 Given a DAG G , a set of variables $R \subset V$, and a value r of R , define the g -functional density

$$\Pr_r(V = v) \equiv f_r(v) \equiv \begin{cases} \prod_{j:V_j \notin R} f(v_j | pa_j) & \text{if } v = (u, r); \\ 0 & \text{if } v = (u, r^*) \text{ with } r^* \neq r. \end{cases}$$

In words, $f_r(v)$ is the density obtained by modifying the product on the right-hand side of (9) by removing the term $f(v_j | pa_j)$ for every $V_j \in R$, while for $V_j \notin R$, for each $R_m \in R$ in PA_j , set R_m to the value r_m in the term $f(v_j | pa_j)$. Note the probability that R equals r is 1 under the density $f_r(v)$, i.e. $\Pr_r(R = r) \equiv f_r(r) = 1$.

The density $f_r(z)$ may not be a well-defined function of the density $f(v)$ of the factual data V when the factual distribution of V is non-positive because setting $R_m \in PA_j$ to the value r_m in $f(v_j | pa_j)$ may result in conditioning on an event that has probability zero of occurring in the factual distribution.

Example 3 In Figure 1 with $R = (X, Z)$, $r = (x = 1, z = 0)$, $f_r(v) \equiv f_{x=1, z=0}(x^*, z^*, y) = f(y | x = 1, z = 0)$ if $(x^*, z^*) = (1, 0)$. On the other hand, $f_{x=1, z=0}(x^*, z^*, y) = 0$ if $(x^*, z^*) \neq (1, 0)$ since, under $f_{x=1, z=0}(x^*, z^*, y)$, X is always 1 and Z always 0. It follows that $f_{x=1, z=0}(y) = f(y | x = 1, z = 0)$. If the event $(X, Z) = (1, 0)$ has probability zero under $f(v)$ then $f_{x=1, z=0}(y)$ is not a function of $f(v)$ and is thus ill-defined.

The following Lemma connects the g -functional density $f_r(v)$ to the intervention density $f_r^{int}(v)$.

Lemma 6 Given an MCM associated with a DAG G , sets of variables $R, Z \subset V$ and a treatment regime r , if the g -functional density $f_r(z)$ is a well-defined function of $f(v)$, then $f_r(z) = f_r^{int}(z)$.

In words, whenever the g -functional density $f_r(z)$ is a well-defined function of $f(v)$, it is equal to the intervention density for Z that would be observed had, contrary to fact, all subjects followed the regime r .

This result can be extended from so-called static treatment regimes r to general treatment regimes, where treatment is a (possibly random) function of the observed history, as follows.

Suppose we are given a set of variables R and for each $V_j = R_m \in R$, we are given a density $p_j(v_j | \bar{v}_{j-1})$. Then we define p_R to be the general treatment regime corresponding to an intervention in which, for each $V_j = R_m \in R$, a subject's treatment level v_j is randomly assigned with randomization probabilities $p_j(v_j | \bar{v}_{j-1})$ that are a function of the values of the subset of the variables \bar{V}_{j-1} that temporally precede V_j . We let $f_{p_R}^{int}(v)$ be the distribution of V that would be observed, if contrary to fact, all subjects had been randomly assigned treatment with probabilities p_R . Further we define the g-functional density $f_{p_R}(v)$ to be the density

$$f_{p_R}(v) \equiv \prod_{j:V_j \notin R} f(v_j | \text{pa}_j) \prod_{j:V_j \in R} p_j(v_j | \bar{v}_{j-1})$$

and for $Z \subset V$, $f_{p_R}(z) \equiv \sum_{v \setminus z} f_{p_R}(v)$. The marginal $f_{p_R}(z)$ is obtained from $f_{p_R}(v)$ by summation in the usual way. Then we have the following extension of Lemma 6.

Extended Lemma 6: *Given an MCM associated with a DAG G , sets of variables $R, Z \subset V$ and a treatment regime p_R , if the g-functional density $f_{p_R}(z)$ is a well-defined function of $f(v)$, then $f_{p_R}(z) = f_{p_R}^{int}(z)$.*

In words, whenever the g-functional density $f_{p_R}(z)$ is a well-defined function of $f(v)$, it is equal to the intervention density for Z that would be observed had, contrary to fact, all subjects followed the general regime p_R . Robins (1986) proved Extended Lemma 6 for a FFRCISTG model; the proof applies equally to a MCM. Extended Lemma 6 actually subsumes Lemma 6 as $f_r^{int}(z)$ is $f_{p_R}^{int}(z)$ for p_R such that for $V_j = R_m \in R$, $p_j(v_j | \bar{v}_{j-1}) = 1$ if $v_j = r_m$ and is zero if $v_j \neq r_m$.

Corollary to Extended Lemma 6: *Given an MCM associated with a DAG G , sets of variables $R, Z \subset V$ and a treatment regime p_R , $f_{p_R}(z) = f_{p_R}^{int}(z)$ whenever p_R satisfies the following positivity condition:*

For all $V_j \in R$, $f(\bar{V}_{j-1}) > 0$ and $p_j(v_j | \bar{V}_{j-1}) \neq 0$ implies

$f(v_j | \bar{V}_{j-1}) \neq 0$ with probability one under $f(v)$.

This follows directly from Extended Lemma 6, as the positivity condition implies $f_{p_R}(z)$ is a well-defined function of $f(v)$. In the literature, one often sees only the Corollary stated

and proved. However, as noted by Gill and Robins (2001), these proofs use the ‘positivity condition’ only to establish that $f_{p_R}(z)$ is a well-defined (i.e. unique) function of $f(v)$. So these proofs are actually proofs of the general version of Extended Lemma 6. In this paper we study models in which $f_{p_R}(z)$ is a well-defined function of $f(v)$ even though the positivity assumption fails; as a consequence, we require the general version of the lemma.

2.6 Agnostic Causal Models

We are now ready to define the agnostic causal model (Spirtes et al., 1993):

Given a DAG G with node set V , the agnostic causal model represented by G assumes that the joint distribution of the factual variables V factors as in (9) and that the interventional density of $Z \subset V$, again denoted by $f_{p_R}^{int}(z)$ or $f_r^{int}(z)$, under treatment regime p_R or regime r is given by the g-functional density $f_{p_R}(z)$ or $f_r(z)$, whenever $f_{p_R}(z)$ or $f_r(z)$ is a well-defined function of $f(v)$.

Although this model assumes that density $f_{p_R}^{int}(v)$ or $f_r^{int}(v)$ of V under these interventions exist, the model makes no reference to counterfactual variables and is agnostic as to their existence. Thus the agnostic causal model does not impose any version of a consistency assumption.

2.7 Interventions restricted to a subset of variables

In this paper we restrict consideration to graphical causal models in which we assume that interventions on every possible subset of the variables are possible and indeed well-defined.

Under any of our four causal models the constraint that a given variable W cannot be intervened upon may be represented in a model by simply requiring that W have no children (hence no descendants) in the DAG. However, this approach may require the inclusion of many additional factual, but unmeasured (and possibly unmeasurable) variables in the DAG. Robins (1986, 1987) present an alternative approach that does not require such an expansion of the set of factual variables. We briefly review this approach in Appendix D. See also the decision-theoretic models of Dawid (2000b) and Heckerman and Shachter (1995).

2.8 Manipulable Contrasts and Parameters

In the introduction we defined the set of manipulable contrasts relative to a graph G to be the set of causal contrasts that are well-defined under the agnostic causal model, i.e., the set of contrasts that are functions of the causal effects $f_{pR}^{int}(z)$. The set consists of all contrasts between treatment regimes in an experiment with sequential treatment assignments, wherein the treatment given at stage m is a function of past covariates on the graph.

Definition 7 *We say a causal effect in a particular causal model associated with a DAG G with node set V is non-parametrically identified from data on V (or equivalently, in the absence of hidden variables) if it is a function of the density $f(v)$ of the factuals V .*

Thus in all four causal models, the causal effects $f_{pR}^{int}(z)$ for which the g-functional $f_{pR}(z)$ is a well-defined function of $f(v)$ are non parametrically identified from data on V . It follows that the manipulable contrasts are non-parametrically identified under an agnostic causal model from observational data with a positive joint distribution and no hidden (i.e. unmeasured) variables. (Recall a discrete joint distribution is positive if the probability of a joint event is nonzero whenever the marginal probability of each individual component of the event is nonzero.)

In contrast, the effect of treatment on the treated

$$\text{ETT}(x) \equiv E[Y(x) - Y(0) \mid X = x]$$

is not a manipulable parameter relative to the graph $G: X \rightarrow Y$; it is not well-defined under the corresponding agnostic causal model. However, $\text{ETT}(x)$ is identified under both MCMs and FFRCISTG models. Robins (2003) stated that a FFRCISTG model only identified “manipulable parameters”. However, in that paper, unlike this one, no explicit definition of manipulable was used; in particular it was not specified which class of interventions was being considered. In the Appendix we show that the MCMs and FFRCISTG models identify $\text{ETT}(x)$, which is not a manipulable parameter relative to the graph $X \rightarrow Y$. However, $\text{ETT}(x)$ is a manipulable parameter relative to an expanded graph G' with deterministic relations; see also Robins et al. (2007) and Geneletti and Dawid (2007).

For expositional simplicity, we will henceforth restrict our discussion to static deterministic regime effects $f_r^{int}(z)$ except when non-static (i.e., dynamic and/or random) regimes p_R are being explicitly discussed.

3 Direct Effects

Consider the query: do cigarettes X have a causal effect on MI Y through a pathway that does not involve hypertension Z ? This query is often rephrased as whether X has a direct causal effect on Y not through the intermediate variable Z . The concept of direct effect has been formalized in three different ways in the literature. For notational simplicity, until the Appendix, we always take X to be binary.

3.1 Controlled Direct Effects (CDE)

Consider a causal model associated with a DAG G with node set V containing (X, Y, Z) . In a counterfactual causal model, the individual and average *controlled direct effect* of X on Y when Z is set to z are respectively defined as $Y(x = 1, z) - Y(x = 0, z)$ and $\text{CDE}(z) = E[Y(x = 1, z) - Y(x = 0, z)]$. In our previous notation, $E[Y(x = 1, z) - Y(x = 0, z)]$ is the difference in means $E_{x=1,z}^{int}[Y] - E_{x=0,z}^{int}[Y]$ of Y under the intervention distributions $f_{x=1,z}^{int}(v)$ and $f_{x=0,z}^{int}(v)$. Under the associated agnostic causal model, counterfactuals do not exist but $\text{CDE}(z)$ can still be defined as $E_{x=1,z}^{int}[Y] - E_{x=0,z}^{int}[Y]$. Under all four causal models, $E_{x=1,z}^{int}[Y] - E_{x=0,z}^{int}[Y]$ is identified from data on V by $E_{x=1,z}[Y] - E_{x=0,z}[Y]$ under the g-formula densities $f_{x=1,z}(v)$ and $f_{x=0,z}(v)$, if these are well-defined functions of $f(v)$. In the case of Figure 1 $E_{x,z}[Y]$ is just the mean $E[Y|X = x, Z = z]$ of the factual Y given $X = x$ and $Z = z$ since, by the definition of the g-formula, $f_{x,z}(y) = f(y | X = x, Z = z)$.

When Z is binary there exist two different controlled direct effects corresponding to $z = 1$ and $z = 0$. For example, $\text{CDE}(1)$ is the average effect of X on Y in the study population were, contrary to fact, all subjects to have Z set to 1. It is possible for $\text{CDE}(1)$ to be zero and $\text{CDE}(0)$ nonzero or vice-versa. Whenever $\text{CDE}(z)$ is nonzero for some level of Z , there will exist a directed path from X to Y not through Z on the causal graph G , regardless of the causal model.

3.2 Pure Direct Effects (PDE)

In a counterfactual model, Robins and Greenland (1992) (hereafter R&G) defined the individual pure direct effect of a (dichotomous) exposure X on Y relative to an intermediate variable Z to be $Y\{x = 1, Z(x = 0)\} - Y(x = 0)$. That is the individual PDE is the subject's value of Y under exposure to X had, possibly contrary to fact, X 's effect on the intermediate Z been blocked (that is, had Z remained at its value under non-exposure) minus the value of Y under non-exposure to X . The individual PDE can also be written as $Y(x = 1, Z(x = 0)) - Y(x = 0, Z(x = 0))$, since $Y(x = 0) = Y(x = 0, Z(x = 0))$. Thus the PDE contrast measures the direct effect of X on Y when Z is set to its value $Z(x = 0)$ under non-exposure to X . The average PDE is given by

$$\begin{aligned} PDE &= E[Y(x = 1, Z(x = 0))] - E[Y(x = 0)] \\ &= E[Y(x = 1, Z(x = 0)) - Y(x = 0, Z(x = 0))]. \end{aligned} \tag{10}$$

Pearl (2001) adopted R&G's definition but changed nomenclature. He refers to the pure direct effect as a 'natural' direct effect. Since the intervention mean $E[Y(x = 0)] = E_{x=0}^{int}[Y]$ is identified from data on V under any of the associated causal models, the PDE is identified if and only if $E[Y\{x = 1, Z(x = 0)\}]$ is identified. Below we will prove that $E[Y(x = 1, Z(x = 0))]$ is not a manipulable effect relative to the graph in Figure 1. Further we show that $E[Y(x = 1, Z(x = 0))]$ is not identified under an MCM or FFRCISTG model from data on V in the absence of further untestable assumptions. However, we shall see that $E[Y(x = 1, Z(x = 0))]$ is identified under the NPSEM associated with the graph in Figure 1.

Under the agnostic causal model, the concept of pure direct effect is not defined since the counterfactual $Y(x = 1, Z(x = 0))$ is not assumed to exist.

3.3 Principal Stratum Direct Effects (PSDE)

In contrast to the control direct effect and pure direct effect, the individual principal stratum direct effect is only defined for subjects for whom X has no causal effect on Z so that $Z(x = 1) = Z(x = 0)$. For a subject with $Z(x = 1) = Z(x = 0) = z$, the individual principal

stratum direct effect is defined to be

$$Y(x = 1, z) - Y(x = 0, z)$$

(here X is assumed to be binary). The average PSDE in principal stratum z is defined to be

$$PSDE(z) \equiv E [Y(1, z) - Y(0, z) \mid Z(1) = Z(0) = z].$$

Robins (1986, Sec. 12.2) first proposed using $PSDE(z)$ to define causal effects. In his paper, $Y = 1$ denoted the indicator of death from a cause of interest (subsequent to a time t), $Z = 0$ denoted the indicator of survival until t from competing causes and the contrast $PSDE(z)$ was used to solve the problem of censoring by competing causes of death in defining the causal effect of the treatment X on the cause Y . Rubin (1998) and Frangakis and Rubin (1999, 2002) later used this same contrast to solve precisely the same problem of ‘censoring by death’. Finally the analysis of Rubin (2004) was also based on this contrast, except that Z and Y were no longer assumed to be failure-time indicators.

The argument given below to prove that $E [Y(x = 1, Z(x = 0))]$ is not a manipulable effect relative to the graph in Figure 1 also proves that $PSDE(z)$ is not a manipulable effect relative to this graph. Furthermore, the $PSDE(z)$ represents a causal contrast on a non-identifiable subset of the study population – the subset with $Z(1) = Z(0) = z$. An even greater potential problem with the PSDE is that if X has an effect on every subject’s Z , then $PSDE(z)$ is undefined for every possible z . If Z is continuous and/or multivariate, it would not be unusual for X to have an effect on every subject’s Z . Thus Z is generally chosen to be univariate and discrete with few levels, often binary when $PSDE(z)$ is the causal contrast.

However, principal stratum direct effects have the potential advantage of remaining well-defined even when controlled direct effects or pure direct effects are ill-defined. Note that for a subject with $Z(x = 1) = Z(x = 0) = z$, we have $Y(x = 1, z) = Y(x = 1, Z(x = 1)) \equiv Y(x = 1)$ and $Y(x = 0, z) = Y(0, Z(0)) \equiv Y(x = 0)$, so the individual PSDE for this subject is $Y(x = 1) - Y(x = 0)$. The average PSDE is given by:

$$PSDE(z) = E [Y(x = 1) - Y(x = 0) \mid Z(1) = Z(0) = z].$$

Thus PSDE’s can be defined in terms of the counterfactuals $Y(x)$ and $Z(x)$. Now in a trial where X is randomly assigned but the intermediate Z is not, there will generally be

reasonable agreement as to the hypothetical intervention (i.e., closest possible world) which sets X to x so $Y(x)$ and $Z(x)$ are well defined; however, there may not be reasonable agreement as to the hypothetical intervention which sets X to x and Z to z , in which case $Y(x, z)$ will be ill-defined. In that event, controlled and pure direct effects are ill-defined, but one can still define $\text{PSDE}(z)$ by the previous display.

However, when $Y(x, z)$ and thus controlled and pure direct effects are ill-defined and therefore use of the $\text{PSDE}(z)$ is proposed, it is often the case that (i) the intermediate variable that is truly of scientific and policy relevance, say Z^* , is many leveled, even continuous and/or multivariate, so $\text{PSDE}(z^*)$ may not exist for any z^* , and (ii) Z is a coarsening (i.e. a function) of Z^* , chosen to ensure that $\text{PSDE}(z)$ exists. In such settings, the counterfactual $Y(x, z^*)$ is frequently meaningful, because the hypothetical intervention which sets X to x and Z^* to z^* (unlike the intervention that sets X to x and Z to z) is well-defined. Furthermore the $\text{CDE}(z^*)$ and PDE based on Z^* , in contrast to the $\text{PSDE}(z)$, provide knowledge of the pathways or mechanisms by which X causes Y and represent the effects of interventions of public health importance. In such a setting, the direct effect contrasts of primary interest are the $\text{CDE}(z^*)$ and the PDE based on Z^* rather than the $\text{PSDE}(z)$ based on a binary coarsening Z of Z^* . See Robins et al. (2007, 2009) for an example and further discussion.

4 Identification of the PDE

We have seen that the $\text{CDE}(z)$, as a manipulable parameter relative to the graph in Figure 1, is generally identified from data on V under all four of the causal models associated with this graph. We next consider identification of $E[Y\{x = 1, Z(x = 0)\}]$ and thus identification of the PDE in three important examples. The first two illustrate that the PDE may be identified in the NPSEM associated with a DAG but not by the associated MCMs or FFRCISTG models. In the third example the PDE is not identified under any of the four causal models associated with the DAG. We will elaborate these examples in subsequent sections.

4.1 Identification of the PDE in the DAG in Fig. 1

Pearl (2001) proved that under the NPSEM associated with the causal DAG in Figure 1 $E[Y\{x = 1, Z(x = 0)\}]$ is identified. To see why, note that if

$$Y(x = 1, z) \perp\!\!\!\perp Z(x = 0) \text{ for all } z, \quad (11)$$

then

$$E[Y(x = 1, Z(x = 0))] = \sum_z E_{x=1,z}^{int}[Y] f_{x=0}^{int}(z), \quad (12)$$

because

$$\begin{aligned} E[Y(x = 1, Z(x = 0))] &= \sum_z E[Y(x = 1, z) | Z(x = 0) = z] pr[Z(x = 0) = z] \\ &= \sum_z E[Y(x = 1, z)] pr[Z(x = 0) = z], \end{aligned}$$

where the first equality is by the laws of probability and the second by (11). Now the right side of Eq. (12) is non-parametrically identified from $f(v)$ under all four causal models since the intervention parameters $E_{x,z}^{int}[Y]$ and $f_x^{int}(z)$ are identified by the g-functional. In particular, with Figure 1 as the causal DAG,

$$\sum_z E_{x=1,z}^{int}[Y] f_{x=0}^{int}(z) = \sum_z E[Y | X = 1, Z = z] f(z | X = 0). \quad (13)$$

Hence it only remains to show that (11) holds for a NPSEM corresponding to the graph in Figure 1. Now we noted in Example 1 that $Y(x = 1, z) \perp\!\!\!\perp Z(x = 0) | X = j$ held for $j = 0$ and $j = 1$ for the NPSEM (but not for the FFRCISTG) associated with the DAG in Figure 1. Further, for this NPSEM, $\{Y(x = 1, z), Z(x = 0)\} \perp\!\!\!\perp X$. Combining, we conclude that (11) holds. In contrast, for a FFRCISTG model or MCM corresponding to Figure 1, $E[Y(x = 1, Z(x = 0))]$ is not identified, because condition (11) need not hold. In Appendix C, we derive sharp bounds for the PDE under the assumption that FFRCISTG model or MCM associated with graph G holds. We find that these bounds may be quite informative, even though the PDE is not (point) identified under this model.

4.2 The ‘Natural Direct Effect’ of Didelez et al. (2006)

Didelez et al. (2006) (DDG) discuss an effect that they refer to as the ‘natural direct effect’ and prove it is identified under the agnostic causal model associated with the DAG in Figure 1, the difference between Eq. (13) and $E[Y | X = 0]$ being the identifying formula. Since the parameter we have referred to as the natural or pure direct effect is not even defined under the agnostic model, it is clear they are giving the same name to a different parameter. Thus DDG’s results have no relevance to the identification of the PDE.

To clarify we discuss DDG’s results in greater detail. To define DDG’s parameter, let $R = (X, Z)$, and consider a regime $p_R \equiv p_{(X=j,Z)}$ with $p(x) = 1$ if and only if $x = j$ and with a given $p(z|x) = p^*(z)$ that does not depend on X . Then $f_{p_{(X=j,Z)}}^{int}(v) = f_{p_{(X=j,Z)}}^{int}(x, z, y)$ is the density in a hypothetical study where each subject receives $X = j$ and then is randomly assigned Z based on the density $p^*(z)$. Then DDG define the natural direct effect to be $E_{p_{(X=1,Z)}}^{int}[Y] - E_{p_{(X=0,Z)}}^{int}[Y]$ with $p^*(z)$ equal to the density $f_{x=0}^{int}(z)$ of Z when X is set to 0, provided $E_{p_{(X=0,Z)}}^{int}[Y]$ is equal to $E_{x=0}^{int}[Y]$, the mean of Y when all subjects are untreated. When $E_{p_{(X=0,Z)}}^{int}[Y] \neq E_{x=0}^{int}[Y]$, they say their natural direct effect is undefined. Now under the agnostic causal model associated with the DAG in Figure 1 it follows from Extended Lemma 6 that $E_{p_{(X=0,Z)}}^{int}[Y] = E_{x=0}^{int}[Y] = E[Y|X = 0]$ and $E_{p_{(X=1,Z)}}^{int}[Y]$ is given by the right side of Eq. (13), confirming DDG’s claim about their parameter $E_{p_{(X=1,Z)}}^{int}[Y] - E_{p_{(X=0,Z)}}^{int}[Y]$. In contrast, our natural or pure direct effect (PDE) parameter is given by the difference between Eq. (13) and $E[Y|X = 0]$ only when $E[Y\{x = 1, Z(x = 0)\}]$ equals Eq. (13), which cannot be the case under an agnostic causal DAG model as $E[Y\{x = 1, Z(x = 0)\}]$ is then undefined. Note $E[Y\{x = 1, Z(x = 0)\}]$ does equal Eq. (13) under the NPSEM associated with Figure 1 but not under the MCM or FFRCISTG model associated with this Figure.

4.3 Identification of the PDE with a measured common cause of Z and Y that is not directly affected by X

Consider the causal DAG in Figure 2(a) that differs from the DAG in Figure 1 in that it assumes (in the context of our smoking study example) there is a measured common cause L of hypertension Z and MI Y that is not caused by X . Suppose we assume an NPSEM

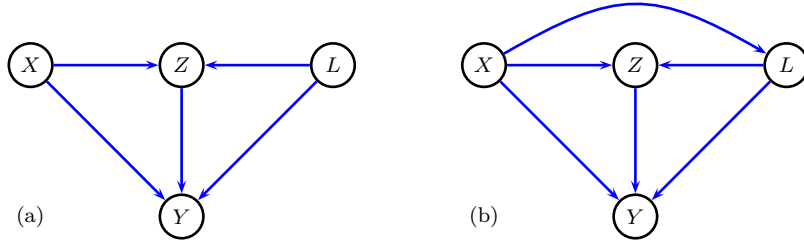


Figure 2: An elaboration of the DAG in Figure 1 in which L is a (measured) common cause of Z and Y .

with $V = (X, L, Z, Y)$ and our goal remains estimation of $E[Y\{x = 1, Z(x = 0)\}]$. Then $E[Y\{x = 1, Z(x = 0)\}]$ remains identified under the NPSEM associated with the DAG in Figure 2(a) with the identifying formula now

$$\sum_{z,l} E[Y | X = 1, Z = z, L = l] f(z | X = 0, L = l) f(l).$$

This follows from the fact that under an NPSEM associated with the DAG in Figure 2(a),

$$Y(x = 1, z) \perp\!\!\!\perp Z(x = 0) | L \quad \text{for all } z, \quad (14)$$

which in turn implies

$$E[Y\{x = 1, Z(x = 0)\}] = \sum_{z,l} E_{x=1,z}^{int}[Y | L = l] f_{x=0}^{int}(z | L = l) f(l). \quad (15)$$

The right side of (15) remains identified under all four causal models via

$$\sum_{z,l} E[Y | X = 1, Z = z, L = l] f(z | X = 0, L = l) f(l). \quad (16)$$

In contrast, for a MCM or a FFRCISTG associated with the graph in Figure 2(a), $E[Y\{x = 1, Z(x = 0)\}]$ is not identified because (14) need not hold.

4.4 Failure of identification of the PDE in an NPSEM with a measured common cause of Z and Y that is directly affected by X

Consider the causal DAG shown in Fig. 2(b) that differs from that in Fig. 2(a) only in that X now causes L so there exists an arrow from X to L . The right side of Eq. (15) remains

identified under all four causal model via

$$\sum_{z,l} E [Y|X = 1, Z = z, L = l] f(z|X = 0, L = l) f(l|X = 0).$$

Under a NPSEM, MCM or FFRCISTG model associated with this causal DAG $Y\{x = 1, Z(x=0)\}$ is by definition

$$Y\{x=1, L(x=1), Z(x=0)\} = Y\{x=1, L(x=1), Z(x=0, L(x=0))\}. \quad (17)$$

Avin et al. (2005) prove that Eq. (14) does not hold for this NPSEM. Thus, even under a NPSEM, we cannot conclude that Eq. (15) holds. In fact, Avin et al. (2005) prove that for this NPSEM $E [Y\{x = 1, Z(x = 0)\}]$ is not identified from data on V . This is because the expression on the RHS of (17) involves both $L(x = 1)$ and $L(x = 0)$, and there is no way to eliminate either.

Additional assumptions identifying the PDE in the NPSEM associated with the DAG in Fig. 2(b)

However, if we were to consider a counterfactual model that imposes even more counterfactual independence assumptions than the NPSEM, then the PDE may still be identified, though by a different formula.

For example, if, in addition to the usual NPSEM independence assumptions, we assume that

$$L(x = 0) \perp\!\!\!\perp L(x = 1) \quad (18)$$

then we have:

$$\begin{aligned}
& E[Y\{x = 1, Z(x = 0)\}] \\
&= \sum_{l^*, l, z} E[Y(x = 1, l, z) \mid L(x = 1) = l, Z(x = 0, l^*) = z, L(x = 0) = l^*] \times \\
&\quad f(L(x = 1) = l, Z(x = 0, l^*) = z, L(x = 0) = l^*) \\
&= \sum_{l^*, l, z} E[Y(x = 1, l, z)] f(L(x = 0) = l^*, L(x = 1) = l) f(Z(x = 0, l^*) = z) \\
&= \sum_{l^*, l, z} E[Y(x = 1, l, z)] f(L(x = 0) = l^*) f(L(x = 1) = l) f(Z(x = 0, l^*) = z) \\
&= \sum_{l^*, l, z} E[Y \mid X = 1, L = l, Z = z] f(L = l^* \mid X = 0) f(L = l \mid X = 1) \times \\
&\quad f(Z = z \mid X = 0, L = l^*). \tag{19}
\end{aligned}$$

Here the second and fourth equalities follows from the usual NPSEM independence restrictions, but the third requires condition (18).

One setting under which (18) holds is that in which the counterfactual variables $L(0)$ and $L(1)$ result from a restrictive ‘minimal sufficient cause model’ (Rothman, 1976) such as:

$$L(x) = (1 - x)A_0 + xA_1 \tag{20}$$

where A_0 and A_1 are independent both of one another, and all other counterfactuals. (Note that (18) would not hold if the right hand side of Eq. (20) was $(1 - x)A_0 + xA_1 + A_2$, even if the A_i ’s were again assumed independent.)

An alternative further assumption, sufficient to identify the PDE in the context of the NPSEM associated with Fig. 2(b), is that $L(1)$ is a deterministic function of $L(0)$, i.e. $L(1) = g(L(0))$ for some function $g(\cdot)$. In this case we have:

$$\begin{aligned}
f(L(x = 0) = l^*, L(x = 1) = l) &= f(L(x = 0) = l^*) I(l = g(l^*)) \\
&= f(L = l^* \mid X = 0) I(l = g(l^*)),
\end{aligned}$$

where $I(\cdot)$ is the indicator function. Hence

$$\begin{aligned}
& E[Y\{x = 1, Z(x = 0)\}] \\
&= \sum_{l^*, l, z} E[Y \mid X = 1, L = l, Z = z] f(L = l^* \mid X = 0) I(l = g(l^*)) \times \\
&\quad f(Z = z \mid X = 0, L = l^*). \tag{21}
\end{aligned}$$

For a scalar L taking values in a continuous state space then there will exist a function $g(\cdot)$ such that $L(1) = g(L(0))$ under the condition of *rank preservation*, i.e. if

$$L_i(0) < L_j(0) \quad \Rightarrow \quad L_i(1) < L_j(1),$$

for all individuals i, j . In this case g is simply the quantile-quantile function:

$$g(l) \equiv F_{L(1)}^{-1}(F_{L(0)}(l)) = F_{L|X=1}^{-1}(F_{L|X=0}(l)), \quad (22)$$

where $F(\cdot)$ and $F^{-1}(\cdot)$ indicate the CDF and its inverse; the equality follows from the NPSEM assumptions; this expression shows that $g(\cdot)$ is identified. (Since L is continuous then the sums over l, l^* in (21) by integrals are replaced by integrals.) A special case of this example is a linear structural equation system, where it was already known that the PDE is identified in the graph in Fig. 2(b). Our analysis shows that identification of the PDE in this graph merely requires rank preservation and not linearity. Note that a linear structural equation model implies both rank preservation and linearity.

We note that the identifying formula (21) differs from by (19). Since neither identifying assumption imposes any restriction on the distribution of the factual variables in the DAG in Figure 2(b), there is no empirical basis for deciding which, if either, of the assumptions are true. Consequently we do not advocate blithely adopting such assumptions in order to preserve identification of the PDE in contexts such as the DAG in Figure 2(b).

5 Models in which the PDE is Manipulable

We now turn to the question of whether $E[Y\{x = 1, Z(x = 0)\}]$ can be identified by intervening on the variables V on G in Figure 1. Now, as noted by R&G (1992) we could observe $E[Y\{x = 1, Z(x = 0)\}]$ if we could intervene and set X to 0, observe $Z(0)$, then “return each subject to their pre-intervention state,” intervene to set X to 1 and Z to $Z(0)$, and finally observe $Y(1, Z(0))$. However, such an intervention strategy will usually not exist because such a return to a pre-intervention state is usually not possible in a real-world intervention (e.g., suppose the outcome Y were death). As a result, because we cannot observe the same subject under both $X = 1$ and $X = 0$, we are unable to directly observe the distribution

of mixed counterfactuals such as $Y\{x = 1, Z(x = 0)\}$. It follows that we cannot observe $E[Y\{x = 1, Z(x = 0)\}]$ by any intervention on the variables X and Z . Pearl (2001) argues similarly. That is, although we can verify through intervention the prediction made by all four causal models that the RHS of Eq. (13) is equal to the expression on the RHS of Eq. (12), we cannot verify, by intervention on X and Z , the NPSEM prediction that Eq. (12) holds.

Thus $E[Y\{x = 1, Z(x = 0)\}]$ is not manipulable with respect to the graph in Figure 1, and hence neither is the PDE with respect to this graph. Yet both of these parameters are identified in the NPSEM associated with this graph. This would be less problematic if these parameters were of little or no substantive interest. However, as shown in the next section, Pearl convincingly argues that such parameters can be of substantive importance.

5.1 Pearl’s substantive motivation for the PDE

Pearl argues that the PDE and the associated quantity $E[Y\{x = 1, Z(x = 0)\}]$ are often causal contrasts of substantive and public health importance by offering examples along the following lines. Suppose a new process can completely remove the nicotine from tobacco, allowing the production of a nicotine-free cigarette to begin next year. The substantive goal is to use already collected data on smoking status X , hypertensive status Z and MI status Y from a randomized smoking cessation trial to estimate the incidence of MI in smokers were all smokers to change to nicotine-free cigarettes. Suppose it is (somehow) known that the entire effect of nicotine on MI is through its effect on hypertensive status, while the non-nicotine toxins in cigarettes have no effect on hypertension. Then, under the further assumption that there do not exist unmeasured confounders for the effect of hypertension on MI, the causal DAG in Figure 1 can be used to represent the study. Under these assumptions, the MI incidence in smokers of cigarettes free of nicotine would be $E[Y\{x = 1, Z(x = 0)\}]$ under all three counterfactual causal models, since the hypertensive status of smokers of nicotine-free cigarettes will equal their hypertensive status under non-exposure to cigarettes. Pearl then assumes a NPSEM and concludes $E[Y\{x = 1, Z(x = 0)\}]$ equals $\sum_z E[Y|X = 1, Z = z] f(z|X = 0)$, and the latter quantity can be estimated from the already available data.

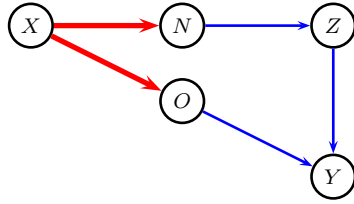


Figure 3: An elaboration of the DAG in Figure 1; N and O are, respectively, the nicotine and non-nicotine components of tobacco.

What is interesting about Pearl’s example is that to argue for the substantive importance of the non-manipulable parameter $E[Y\{x = 1, Z(x = 0)\}]$, he tells a story about the effect of a manipulation – a manipulation that makes no reference to Z at all. Rather, the manipulation is to intervene to eliminate the nicotine component of cigarettes.

Indeed, the most direct representation of his story is provided by the extended DAG in Figure 3 with $V = (X, N, O, Z, Y)$ where N is a binary variable representing nicotine exposure, O is a binary variable representing exposure to the non-nicotine components of a cigarette, and (X, Z, Y) are as above. The bolded arrows from X to N and O indicate a deterministic relationship. Specifically in the factual data, with probability one under $f(v)$, either one smokes normal cigarettes so $X = N = O = 1$ or one is a nonsmoker (i.e. ex-smoker) and $X = N = O = 0$. In this representation the parameter of interest is the mean $E_{n=0,o=1}^{int}[Y]$ of Y had, contrary to fact, all subjects only been exposed to the non-nicotine components. As $E_{n=0,o=1}^{int}[Y]$ is a function of $f_{n=0,o=1}^{int}(v)$, we conclude that $E_{n=0,o=1}^{int}[Y]$ is a manipulable causal effect relative to the DAG in Figure 3. Further, Pearl’s story gives no reason to believe that there is any confounding for estimating this effect. In the Appendix we present a scenario that differs from Pearl’s in which $E_{n=0,o=1}^{int}[Y]$ is confounded, and thus none of the four causal models associated with Figure 3 can be true (even though the FFRICISTG and agnostic causal models associated with Figure 1 are true).

In contrast, under Pearl’s scenario it is reasonable to take any of the four causal models, including the agnostic model, associated with Figure 3 as true. Under such a supposition $E_{n=0,o=1}^{int}[Y]$ is identified if $E_{n=0,o=1}[Y]$ is a well-defined function of $f(v)$. Note, under $f(v)$, data on (X, Z, Y) is equivalent to data on $V = (X, N, O, Z, Y)$, since X completely determines O and N in the factual data. We now show that, with Figure 3 as the causal DAG

and $V = (X, N, O, Z, Y)$, under all four causal models, $E_{n=0,o=1}^{int}[Y]$ is identified simply by applying the g-formula density in standard fashion. This result may seem surprising at first since no subject in the actual study data followed the regime ($n = 0, o = 1$), so the standard positivity assumption $\Pr_V[N = 0, O = 1] > 0$ usually needed to make the g-formula density $f_{n=0,o=1}(v)$ a function of $f(v)$ [and thus identifiable] fails.

However, as we now demonstrate, even without positivity, the conditional independences implied by the assumptions of no direct effect of N on Y and no effect of O on Z encoded in the missing arrows from N to Y and O to Z in Figure 3 along with the deterministic relationship between O, N , and X under $f(v)$ allow one to obtain identification. Specifically, under the DAG in Figure 3,

$$\begin{aligned} f_{n=0,o=1}(y, z) &= f(y \mid O = 1, z)f(z \mid N = 0) \\ &= f(y \mid O = 1, N = 1, z)f(z \mid N = 0, O = 0) \\ &= f(y \mid X = 1, z)f(z \mid X = 0), \end{aligned}$$

where the first equality is by definition of the g-formula density $f_{n=0,o=1}(y, z)$, the second by the conditional independence relations encoded in the DAG in Figure 3, and the last by the deterministic relationships between O, N , and X under $f(v)$ with $V = (X, N, O, Z, Y)$.

Thus

$$\begin{aligned} E_{n=0,o=1}[Y] &\equiv \sum_{y,z} y f_{n=0,o=1}(y, z) \\ &= \sum_{y,z} y f_V(y \mid X = 1, z) f(z \mid X = 0) \\ &\equiv \sum_z E[Y \mid X = 1, Z = z] f(z \mid X = 0), \end{aligned}$$

which is a function of $f(v)$ with $V = (X, N, O, Z, Y)$. Note that this argument goes through even if Z and/or Y are non-binary, continuous variables.

The rôle of the extended causal model in Figure 3

The identifying formula under all four causal models associated with the DAG in Figure 3 is the identifying formula Pearl obtained when representing the problem as the estimation of $E[Y\{x = 1, Z(x = 0)\}]$ under the NPSEM associated with the DAG in Figure 1.

For Pearl, having at the outset assumed an NPSEM associated with the DAG in Figure 1, the story did not contribute to identification; rather it only served to show that the non-manipulable parameter $E[Y\{x = 1, Z(x = 0)\}]$ of the NPSEM associated with the DAG in Figure 1 could, under the scenario of our story, encode a substantively important parameter – the manipulable causal effect of setting N to 0 and O to 1 on the extended causal model associated with the DAG in Figure 3. However, from the Popperian point of view, it is the story itself that makes Pearl’s claim that $E[Y\{x = 1, Z(x = 0)\}] = \sum_z E[Y | X = 1, z] f(z | O = 1)$ refutable and thus scientifically meaningful. Specifically, when nicotine free cigarettes become available, Pearl’s claim can be tested by an intervention that forces a random sample of the population to smoke nicotine free cigarettes.

For someone only willing to entertain an agnostic causal model, the information necessary to identify the effect of nicotine-free cigarettes was contained in the story as the parameter $E[Y\{x = 1, Z(x = 0)\}]$ is undefined without the story. [Someone, such as Dawid (2000b), opposed to counterfactuals and thus wedded to the agnostic causal model might then reasonably and appropriately choose to define $E_{n=0,o=1}^{int}[Y] - E_{n=0,o=0}^{int}[Y] = E_{n=0,o=1}^{int}[Y] - E_{x=0}^{int}[Y]$ to be the natural or pure direct effect of X not through Z . This definition differs from, and in our view is preferable to, the definition of DDG (2006) discussed in §4.2: the definition of DDG fails to correspond to the concept of the PDE as used in the literature since its introduction in Robins and Greenland (1992).]

For an analyst who had only assumed the MCM, but not necessarily the NPSEM, associated with the DAG in Figure 1 was true, the information contained in the above story licenses the assumption that the MCM associated with Figure 3 holds. This latter assumption can be used in two alternative ways, both leading to the same identifying formula. First, it leads via Lemma 6 to the above g-functional analysis also used by the agnostic model advocate. Second, as we next show, it can be used to prove that (11) holds, allowing identification to proceed a la Pearl (2001).

The rôle of determinism

Consider an MCM associated with the DAG in Figure 3 with node set $V = (X, N, O, Z, Y)$. It follows from the fact that $X = N = O$ w.p. 1, that the condition that $N(x) = O(x) = x$ w.p. 1

also holds. However, for pedagogic purposes, suppose for the moment that the condition $N(x) = O(x) = x$ w.p. 1 does not hold.

For expositional simplicity we assume all variables are binary so our model is also an FFRCISTG model. Then, $V_0=X$, $V_1(v_0) = N(x)$, $V_2(v_0, v_1) = V_2(v_1) = O(x)$, $V_3(\bar{v}_2) = V_3(v_1) = Z(n)$, and $V_4(\bar{v}_3) = V_4(v_2, v_3)$. By Theorem 1, $\{Y(o, z), Z(n), O(x), N(x)\}$ are mutually independent. However, because we are assuming an FFRCISTG model and not an NPSEM, we cannot conclude that $O(x) \perp\!\!\!\perp N(x^*)$ for $x \neq x^*$.

Consider the induced counterfactual models for the variables (X, Z, Y) obtained from our FFRCISTG model by marginalizing over (N, O) . Because N and O each have only a single child on the graph in Figure 3, the counterfactual model over (X, Z, Y) is the FFRCISTG associated with the complete graph of Figure 1, where the one step ahead counterfactuals $Z^{(1)}(x), Y^{(1)}(x, z)$ associated with Figure 1 are obtained from the counterfactuals $\{Y(o, z), Z(n), O(x), N(x)\}$ associated with Figure 3 by $Z^{(1)}(x) = Z(N(x))$, $Y^{(1)}(x, z) = Y(O(x), z)$. Here we have used the superscript ‘(1)’ to emphasize the graph with respect to which $Z^{(1)}(x)$ and $Y^{(1)}(x, z)$ are one step ahead counterfactuals. We cannot conclude that $Z^{(1)}(0) = Z(N(0))$ and $Y^{(1)}(1, z) = Y(O(1), z)$ are independent, even though $Z(n)$ and $Y(o, z)$ are independent because, as noted above, the FFRCISTG model associated with Figure 3 does not imply independence of $O(1)$ and $N(0)$.

Suppose now we re-instate the deterministic constraint that $N(x) = O(x) = x$ w.p. 1. Then we conclude $O(x)$ is independent of $N(x^*)$, since both variables are constants. It then follows that $Z^{(1)}(0)$ and $Y^{(1)}(1, z)$ are independent and thus that Eq. (11) is true and $E[Y^{(1)}(1, Z^{(1)}(0))]$ is identified.

The need for conditioning on events of probability zero

In our argument that, under the deterministic constraint that $N(x) = O(x) = x$ w.p. 1, the FFRCISTG associated with the DAG in Figure 3 implied condition (11), the crucial step was the following: by Theorem 1, the independencies in Eq. (1) that define an FFRCISTG imply that $Y(o, z)$ and $Z(n)$ are independent for $n = 0$ and $o = 1$. In this section, we show that had we modified Eq. (1) and thus our definition of an FFRCISTG by restricting to conditioning events $\bar{V}_{m-1} = \bar{v}_{m-1}$ that have a positive probability under $f(v)$, then Theorem 1 would not

hold for non-positive densities $f(v)$. Specifically, if $f(v)$ is not positive, the modified version of Eq. (1) does not imply Eq. (6); furthermore the set of independencies implied by a modified FFRCISTG associated with a graph G could differ for different orderings of the variables consistent with the descendant relationships on the graph. Specifically, we now show that for the modified FFRCISTG associated with Figure 3 and the ordering (X, N, O, Z, Y) , we cannot conclude $Y(x, n, o, z) = Y(o, z)$ and $Z(x, n, o) = Z(n)$ are independent for $n = 0$ and $o = 1$ and thus that Eq. (11) holds. However the modified FFRCISTG with the alternative ordering (X, N, Z, O, Y) does imply $Y(o, z) \perp\!\!\!\perp Z(n)$. First consider the modified FFRCISTG associated with Figure 3 and ordering (X, N, O, Z, Y) under the deterministic constraint $N(x) = O(x) = x$ w.p. 1. The unmodified Eq. (1) implies the set of independencies

$$Y\{n, o, z\} \perp\!\!\!\perp Z(n, o) \mid X = x, N(x) = n, O(x) = o, \text{ for } \{z, x, n, o \in \{0, 1\}\}.$$

The modified Eq. (1) only implies the subset corresponding to $\{x, z \in \{0, 1\}; n = o = x\}$ since the event $\{N(x) = j, O(x) = 1 - j, j \in \{0, 1\}\}$ has probability zero. As a consequence, we can only conclude that $Y\{n, o, z\} = Y\{o, z\} \perp\!\!\!\perp Z(n)$ for $o = n$.

In contrast, for the modified FFRCISTG associated with Figure 3 and the ordering $V = (X, N, Z, O, Y)$, the deterministic constraint implies $N(x) = O(x) = x$ w.p. 1 implies $Y(o, z) \perp\!\!\!\perp Z(n)$ for $n = 0$ and $o = 1$ as follows: By Eq. (1) and the fact that $Y\{x, n, z, o\} = Y(o, z)$ and $Z(x, n) = Z(n)$, we have, without having to condition on an event of probability zero, that

$$\{Y(o, z), Z(n)\} \perp\!\!\!\perp X \text{ for } z, o, n \in \{0, 1\}, \quad (23)$$

$$Y(o, z) \perp\!\!\!\perp Z(n) \mid X = x, N(x) = n \text{ for } x, z, o \in \{0, 1\} \text{ and } n = x. \quad (24)$$

But (23) implies $Y(o, z) \perp\!\!\!\perp Z(n = x) \mid X = x$ for $x, z, o \in \{0, 1\}$ as $X = x$ is the same event as $X = N(x) = x$. Thus $Y(o, z) \perp\!\!\!\perp Z(n)$ for $n, z, o \in \{0, 1\}$ by (23).

The heuristic reason that for the ordering $V = (X, N, Z, O, Y)$ we must condition on events of probability zero in Eq. (1) in order to prove (11) is that such conditioning is needed to instantiate the assumption that O has no effect on Z ; if we do not allow conditioning on events of probability zero, the FFRCISTG model with this ordering does not instantiate this assumption because O and N are equal with probability one and thus we can substitute O

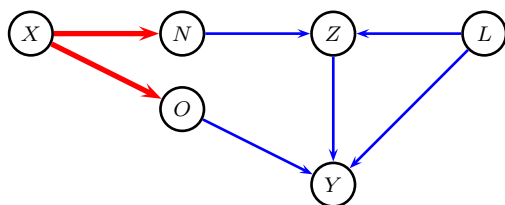


Figure 4: The graph from Figure 3 with, in addition, a measured common cause of the intermediate Z and the final response Y .

for N as the cause Z . Under the ordering $V = (X, N, Z, O, Y)$ in which O is subsequent to Z , it was not necessary to condition on events of probability zero in Eq. (1) to instantiate this assumption, as the model precludes later variables in the ordering from being causes of earlier variables; thus O cannot be a cause of Z .

The above example demonstrates that the assumption that Eq. (1) holds even when we condition on events of probability zero can place independence restrictions on the distribution of the counterfactuals over and above those implied by the assumption that Eq. (1) holds when the conditioning events have positive probability. One might wonder how this could be so; it is usually thought that different choices for probabilities conditional on events of probability zero have no distributional implications. The following simple canonical example that makes no reference to causality or counterfactuals clarifies how multiple distributional assumptions conditional on events of probability zero can place substantive restrictions on a distribution.

Example 4 Suppose we have random variables (X, Y, R) where $R = 1$ with probability one. Suppose we assume both that (i) $f(x, y | R = 0) = f(x, y)$ and (ii) $f(x, y | R = 0) = f(x | R = 0) f(y | R = 0)$. Then we can conclude that $f(x, y) = f(x | R = 0) f(y | R = 0)$ and thus that X and Y are independent since the joint density $f(x, y)$ factors as a function of x times a function of y . The point is that although neither assumption (i) nor assumption (ii) alone restrict the joint distribution of (X, Y) , nonetheless, together they impose the restriction that X and Y are independent.

Inclusion of a measured common cause of Z and Y

A similar elaboration may be given for the causal DAG in Figure 2(a). The extended causal DAG represented by our story would then be the DAG in Figure 4. Under any of our four causal models,

$$\begin{aligned} f_{n=0,o=1}(y, z, l) &= f(y \mid O = 1, z, l)f(z \mid N = 0, l)f(l) \\ &= f(y \mid O = 1, N = 1, z, l)f(z \mid N = 0, O = 0, l)f(l) \\ &= f(y \mid X = 1, z, l)f(z \mid X = 0, l)f(l). \end{aligned}$$

Hence,

$$E_{n=0,o=1}[Y] = \sum_{z,l} E[Y \mid X = 1, Z = z, L = l] f(z \mid X = 0, L = l)f(l),$$

which is the identifying formula Pearl obtained when representing the problem as the estimation of $E[Y\{x = 1, Z(x = 0)\}]$ under an NPSEM associated with the DAG in Figure 2(a).

Summary

We believe in some generality that whenever a particular causal effect is (a) identified from data on V under a NPSEM associated with a DAG G with node set V (but is not identified under the associated MCM, FFRCISTG model or agnostic causal model) and (b) can be expressed as the effect of an intervention on certain variables (which may not be elements of V) in an identifiable sub-population, then that causal effect is also identified under the agnostic causal DAG model based on a DAG G' with node set V' , a superset of V . To find such an identifying causal DAG model G' , it is generally necessary to make the variables in $V' \setminus V$ deterministic functions of the variables in V . The above examples based on extended DAGs in Figures 3 and 4 are cases in point; see Geneletti and Dawid (2007); Robins et al. (2007) and Appendix A for such a construction for the effect of treatment on the treated.

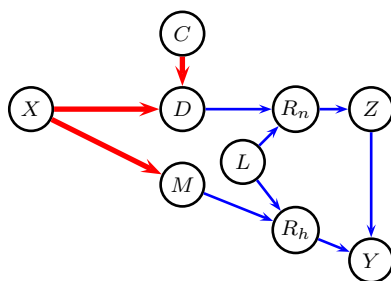


Figure 5: An example in which an interventional interpretation of the PDE is hard to conceive.

5.2 An example in which an interventional interpretation of the PDE is more controversial

The following example shows that the construction of a scientifically plausible story under which the PDE can be regarded as a manipulable contrast relative to an expanded graph G' may be more controversial than our previous example would suggest. After presenting the example we briefly discuss its philosophical implications.

Suppose nicotine X was the only chemical found in cigarettes that had an effect on MI, but nicotine produced its effects by two different mechanisms. First, it increased blood pressure Z by directly interacting with a membrane receptor on blood pressure control cells located in the carotid artery in the neck. Second it directly caused atherosclerotic plaque formation and thus an MI by directly interacting with a membrane receptor of the same type located on the endothelial cells of the coronary arteries of the heart. Suppose the natural endogenous ligand produced by the body that binds to these receptors was nicotine itself. Finally assume that exogenous nicotine from cigarettes had no causal effect on the levels of endogenous nicotine (say, because the time-scale under study is too short for homeostatic feedback mechanisms to kick in) and we had precisely measured levels of endogenous nicotine L before randomizing to smoking or not smoking (X). Suppose that based on this story, an analyst posits the NPSEM associated with the graph in Figure 2(a) with $V = (X, Z, Y, L)$ is true. As noted in §4.3, under this supposition $E[Y\{x = 1, Z(x = 0)\}]$ is identified via $\sum_{z,l} E[Y | X = 1, Z = z, L = l] f(z | X = 0, L = l) f(l)$.

Can we express $E[Y\{x = 1, Z(x = 0)\}]$ as an effect of a scientifically plausible interven-

tion? To do so, we must devise an intervention that (i) blocks the effect of exogenous nicotine on the receptors in the neck without blocking the effect of exogenous nicotine on the receptors in the heart, but (ii) does not block the effect of endogenous nicotine on the receptors in either the neck or heart. To accomplish (i), one could leverage the physical separation of the heart and the neck to build a “nano-cage” around the blood pressure control cells in the neck that prevents exogenous nicotine from reaching the receptors on these cells. However, because endogenous and exogenous nicotine are chemically and physically identical, the cage would also block the effect of endogenous nicotine on receptors in the neck, in violation of (ii). Thus a critic might conclude that $E[Y\{x = 1, Z(x = 0)\}]$ could not be expressed as the effect of an intervention. If the critic adhered to the slogan “no causation without manipulation” (i.e. causal contrasts are best thought of in terms of explicit interventions that, at least in principle, could be performed (Robins and Greenland, 2000)) he or she would then reject the PDE as a meaningful causal contrast in this context. In contrast, if the critic believed in the ontological primacy of causation, they would take the example as evidence for their slogan “causation before manipulation”.

Alternatively, one can argue that the critic’s conclusion that $E[Y\{x = 1, Z(x = 0)\}]$ could not be expressed as the effect of an intervention only indicates a lack of imagination and an intervention satisfying (i) and (ii) may someday exist. Specifically, some day it may be possible to chemically attach a side group to the exogenous nicotine in cigarettes in such a way that a) the effect of the (exogenous) chemically-modified nicotine and the effect of the unmodified nicotine on the receptors in the heart and neck are identical, while b) allowing the placement of a “nano-cage” in the neck that successfully binds the side group attached to the exogenous nicotine, thereby preventing it from reaching the receptors in the neck. In that case $E[Y\{x = 1, Z(x = 0)\}]$ equals a manipulable contrast of the extended deterministic causal DAG of Figure 5. In the Figure $C = 1$ denotes the “nano-cage” is present. We allow X to take three values, as before $X = 0$ indicates no cigarette exposure, $X = 1$ indicates exposure to cigarettes with unmodified nicotine, and $X = 2$ indicates exposure to cigarettes with modified nicotine. R_n is the fraction of the receptors in the neck that are bound to a nicotine molecule (exogenous *or* endogenous), R_h is the fraction of the receptors in the heart that are bound to a nicotine molecule. M is a variable that is 1 if and only if $X \neq 0$;

D is a variable that takes the value 1 if and only if either $X = 1$ or ($X = 2$ and $C = 0$). Then $E[Y\{x = 1, Z(x = 0)\}]$ is the parameter $E_{x=2, c=1}^{int}[Y]$ corresponding to the intervention described in a) and b). Under all four causal models associated with the graph in Figure 5,

$$\begin{aligned} f_{x=2, c=1}[y, z, l] &\equiv \sum_{m, d, r_h, r_n} f[y | r_h, z] f[r_h | m, l] f[z | r_n, l] f(m | x = 2) \\ &\quad \times f(r_n | d, l) f(d | c = 1, x = 2) f(l) \\ &= f[y | M = 1, l, z] f[z | D = 0, l] f(l) \\ &= f[y | X = 1, z, l] f[z | X = 0, l] f(l), \end{aligned}$$

where the first equality uses the fact that $D = 0$ and $M = 1$ when $x = 2$ and $c = 1$ and the second uses the fact that, since in the observed data $C = 0$ w.p. 1, $D = 0$ if and only if $X = 0$, and $M = 1$ if and only if $X = 1$ (since $X \neq 2$ w.p. 1). Thus

$$E_{x=2, c=1}[Y] = \sum_{z, l} E[Y | X = 1, Z = z, L = l] f(z | X = 0, L = l) f(l),$$

which is the identifying formula Pearl obtained when representing the problem as the estimation of $E[Y\{x = 1, Z(x = 0)\}]$ under an NPSEM based on the DAG in Figure 2(a).

As noted in the introduction, the exercise of trying to construct a story to provide an interventionist interpretation for a non-manipulable causal parameter of a NPSEM often helps one devise explicit, and sometimes even practical, interventions which can then be represented as a manipulable causal effect relative to an extended deterministic causal DAG model such as Figure 3.

6 Path Specific Effects

In this section we extend our results to path specific effects. We begin with a particular motivating example.

6.1 A Specific Example

Suppose our underlying causal DAG was the causal DAG of Figure 2(b) in which there is an arrow from X to L . We noted above that Pearl proved $E[Y\{x = 1, Z(x = 0)\}]$ was

not identified from data (X, L, Z, Y) on the causal DAG in Figure 2(b) even under the associated NPSEM. There exist exactly three possible extensions of Pearl's original story that are consistent with the causal DAG in Figure 2(b), as shown in Figure 6: (a) nicotine N causes L but O does not; (b) O causes L but N does not; (c) both N and O cause L . We consider as before the causal effect $E_{n=0,o=1}^{int}[Y]$. Under all four causal models associated with graph (a) in Figure 6, $E_{n=0,o=1}^{int}[Y]$ is identified from factual data on $V = (X, L, Z, Y)$. Specifically, on the DAG in Figure 6(a), we have

$$\begin{aligned}
f_{n=0,o=1}(y, z, l) &= f(y \mid O = 1, z, l)f(z \mid N = 0, l)f(l \mid N = 0) \\
&= f(y \mid O = 1, N = 1, z, l)f(z \mid N = 0, O = 0, l)f(l \mid N = 0, O = 0) \\
&= f(y \mid X = 1, z, l)f(z \mid X = 0, l)f(l \mid X = 0),
\end{aligned}$$

so

$$E_{n=0,o=1}^{int}[Y] = \sum_{l,z} E(Y \mid X = 1, z, l)f(z \mid X = 0, l)f(l \mid X = 0). \quad (25)$$

Similarly, under all four causal models associated with graph (b) in Figure 6, $E_{n=0,o=1}^{int}[Y]$ is identified from factual data on $V = (X, L, Z, Y)$: on the DAG in Figure 6(b) we have

$$\begin{aligned}
f_{n=0,o=1}(y, z, l) &= f(y \mid O = 1, z, l)f(z \mid N = 0, l)f(l \mid O = 1) \\
&= f(y \mid X = 1, z, l)f(z \mid X = 0, l)f(l \mid X = 1)
\end{aligned}$$

so

$$E_{n=0,o=1}^{int}[Y] = \sum_{l,z} E(Y \mid X = 1, z, l)f(z \mid X = 0, l)f(l \mid X = 1). \quad (26)$$

However, $E_{n=0,o=1}^{int}[Y]$ is not identified from factual data on $V = (X, L, Z, Y)$ under any of the four causal models associated with graph (c) in Figure 6. In this graph $f_{n=0,o=1}(y, z, l) = f(y \mid O = 1, z, l)f(z \mid N = 0, l)f(l \mid O = 1, N = 0)$. However, $f(l \mid O = 1, N = 0)$ is not identified from the factual data since the event $\{O = 1, N = 0\}$ has probability zero under $f(v)$. Note that the identifying formulae for $E_{n=0,o=1}^{int}[Y]$ for the graphs in Figure 6 (a) and (b) are different.

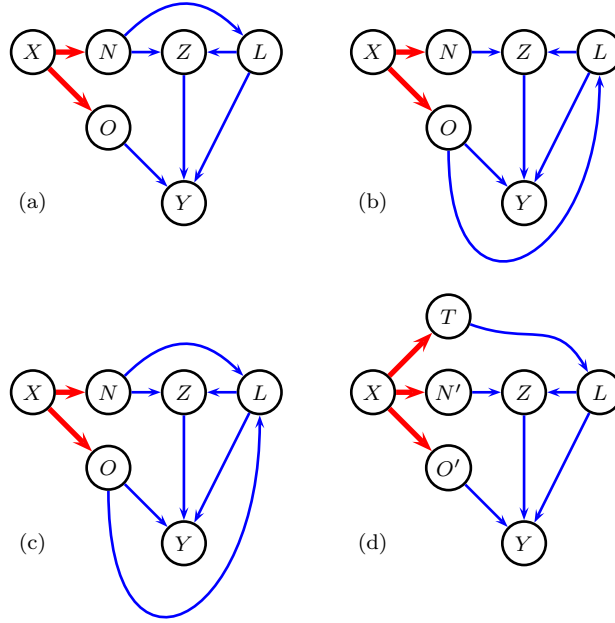


Figure 6: Elaborations of the graph in Figure 2(b), with additional variables as described in the text.

Relation to counterfactuals associated with the DAG in Fig. 2(b)

Let $Y(x, l, z)$, $Z(x, l)$ and $L(x)$ denote the one step ahead counterfactuals associated with the graph in Figure 2(b). Then, it is clear from the assumed deterministic counterfactual relation $N(x) = O(x) = x$, that the parameter

$$E_{n=0, o=1}^{int} [Y] = E [Y\{o = 1, L(n = 0), Z(n = 0, L(n = 0))\}]$$

associated with the graph in Figure 6(a) can be written in terms of the counterfactuals associated with the graph in Figure 2(b) as

$$E [Y\{x = 1, L(x = 0), Z(x = 0)\}] = E [Y\{x = 1, L(x = 0), Z(x = 0, L(x = 0))\}].$$

Likewise, we have that the parameter

$$E_{n=0, o=1}^{int} [Y] = E [Y\{o = 1, L(o = 1), Z(n = 0, L(o = 1))\}]$$

associated with the graph in Figure 6(b) equals

$$E [Y\{x = 1, L(x = 1), Z(x = 0, L(x = 1))\}]$$

in terms of the counterfactuals associated with the graph in Figure 2(b). In contrast $E_{n=0, o=1}^{int} [Y]$ associated with the graph in Figure 6(c) is not the mean of any counterfactual defined from $Y(x, l, z)$, $Z(x, l)$ and $L(x)$ under the graph in Figure 2(b) since L , after

intervening to set $n = 0$, $o = 1$, is neither $L(x = 1)$ nor $L(x = 0)$, since both imply a counterfactual for L under which $n = o$.

Furthermore, the parameter

$$E[Y(x = 1, Z(x = 0))] = E[Y\{x = 1, L(x = 1), Z(x = 0, L(x = 0))\}]$$

associated with the graph in Figure 2(b) is not identified under any of the four causal models associated with any of the three graphs in Fig. 6(a), (b) and (c); see §4.4.

Thus, in summary, under an MCM or FFRCISTG model associated with the DAG in Fig. 6(a), the extension of Pearl's original story encoded in that DAG allows the identification of the causal effect $E[Y\{x = 1, L(x = 0), Z(x = 0)\}]$ associated with the DAG in Fig. 2(b). Similarly, under an MCM or FFRCISTG model associated with the DAG in Fig. 6(b) the extension of Pearl's original story encoded in this graph allows the identification of the causal effect $E[Y\{x = 1, L(x = 1), Z(x = 0, L(x = 1))\}]$ associated with the DAG in Fig. 2(b).

Contrast with the NPSEM for the DAG in Figure 2(b)

We now compare these results to those obtained under the assumption that the NPSEM associated with the DAG in Fig. 2(b) held. Under this model Avin et al. (2005) proved using their theory of path-specific effects that while $E[Y\{x = 1, Z(x = 0)\}]$ is unidentified, both

$$E[Y(x = 1, L(x = 0), Z(x = 0))] \text{ and } E[Y\{x = 1, L(x = 1), Z(x = 0, L(x = 1))\}] \quad (27)$$

are identified (without requiring any additional story) by Equations (25) and (26) respectively.

From the perspective of the FFRCISTG models associated with the graphs in Figure 6(a) and (b) if N and O represent, as we have been assuming, the substantive variables Nicotine and Other components of cigarettes (rather than merely formal mathematical constructions), these graphs will generally represent mutually exclusive causal hypotheses. As a consequence at most one of the two FFRCISTG models will be true; thus, from this perspective, only one of the two parameters of (27) will be identified.

Simultaneous identification of both parameters in (27) by an expanded graph

We next describe an alternative scenario associated with the expanded graph in Fig. 6(d) whose substantive assumptions imply (i) the FFRCISTG model associated with Fig. 6(d) holds and (ii) the two parameters of (27) are manipulable parameters of that FFRCISTG which are identified by Equations (25) and (26), respectively. Thus, this alternative scenario provides a (simultaneous) manipulative interpretation for the non-manipulative (relative to (X, Z, Y)) parameters (27) that are simultaneously identified by the NPSEM associated with the DAG in Fig. 2(b).

Suppose it was (somehow?) known that, as encoded in the DAG in Fig. 6(d), the Nicotine (N') component of cigarettes was the only (cigarette-related) direct cause of Z not through L , the Tar (T) component was the only (cigarette-related) direct cause of L , the other components (O') contained all the (cigarette-related) direct causes of Y not through Z and L , and there are no further confounding variables so that the FFRCISTG model associated with Fig. 6(d) can be assumed true. Then the parameter $E_{n'=0, t=0, o'=1} [Y]$ associated with Fig. 6(d) equals both the parameter $E [Y(x=1, L(x=0), Z(x=0))]$ associated with Fig. 2(b) and the parameter $E_{n=0, o=1} [Y]$ associated with Fig. 6(a) (where $n=0$ is now defined to be the intervention that sets nicotine $n' = 0$ and tar $t = 0$ while $o = 1$ is the intervention $o' = 1$; N and O are redefined by $1 - N = (1 - N')(1 - T)$ and $O = O'$). Furthermore, $E_{n'=0, t=0, o'=1} [Y]$ is identified by Equation (25). Similarly, the parameter $E_{n'=0, t=1, o'=1} [Y]$ associated with Fig. 6(d) is equal to both the parameter $E[Y\{x=1, L(x=1), Z(x=0, L(x=1))\}]$ associated with Fig. 2(b) and the parameter $E_{n=0, o=1} [Y]$ associated with Fig. 6(b) (where $n = 0$ is now the intervention that sets nicotine $n' = 0$ while $o = 1$ denotes the intervention that sets tar $t = 1$ and $o' = 1$; N and O are redefined by $N = N'$ and $O = TO'$). Furthermore, the parameter $E_{n'=0, t=1, o'=1} [Y]$ is identified by Equation (26).

Note, under this alternative scenario, and in contrast to our previous scenarios, the substantive meanings of the intervention that sets $n = 0$ and $o = 1$ and of the variables N and O for Fig. 6(a) differs from the substantive meaning of this intervention and these variables for Fig. 6(b), allowing the two parameters $E_{n=0, o=1} [Y]$ to be identified simultaneously, each by a different formula, under the single FFRCISTG model associated with Fig. 6(d).

Connection to Path Specific Effects

Avin et al. (2005) refer to $E[Y(x = 1, L(x = 0), Z(x = 0))]$ as the effect of $X = 1$ on Y when the paths from X to L and from X to Z are both blocked (inactivated) and to $E[Y\{x = 1, L(x = 1), Z(x = 0, L(x = 1))\}]$ as the effect of $X = 1$ on Y when the paths from X to Z are blocked. They refer to

$$E[Y\{x = 1, Z(x = 0)\}] = E[Y\{x = 1, L(x = 1), Z(x = 0, L(x = 0))\}]$$

as the effect of $X = 1$ on Y when both the path from X to Z and (X 's effect on) the path from L to Z is blocked.

6.2 The General Case

We now generalize the above results. Specifically, given any DAG G , with a variable X , construct a deterministic extended DAG G_{ex} that differs from G only in that the only arrows out of X on G_{ex} are deterministic arrows from X to new variables N and O and the origin of each arrow out of X on G is from either N or O (but never both) on G_{ex} . Then, with $V \setminus X$ being the set of variables on G other than X , the marginal g-formula density $f_{n=0,o=1}(v \setminus x)$ is identified from the distribution of the variables V on G whenever $f(v)$ is a positive distribution by

$$\begin{aligned} f_{n=0,o=1}(v \setminus x) = & \prod_{\{j; V_j \text{ is not a child of } X \text{ on } G\}} f(v_j \mid pa_j) \times \\ & \prod_{\{j; V_j \text{ is a child of } O \text{ on } G_{ex}\}} f(v_j \mid pa_j \setminus x, X = 1) \times \\ & \prod_{\{j; V_j \text{ is a child of } N \text{ on } G_{ex}\}} f(v_j \mid pa_j \setminus x, X = 0). \end{aligned}$$

Note if X has p children on G , there exist 2^p different graphs G_{ex} . The identifying formula for $f_{n=0,o=1}(v \setminus x)$ in terms of $f(v)$ depends on the graph G_{ex} . It follows that, under the assumption that a particular G_{ex} is associated with one of our four causal models, the intervention distribution $f_{n=0,o=1}^{int}(v \setminus x)$ corresponding to that G_{ex} is identified under any of the four associated models.

We now discuss the relationship with path-specific effects. Avin et al. (2005) first define, for any counterfactual model associated with G , the path specific effect on the density of

$V \setminus X$ when various paths on graph G have been blocked. Avin et al. (2005) further determine which path specific densities are identified under the assumption that the NPSEM associated with G is true, and provide the identifying formulae.

The results of Avin et al. (2005) imply that the path-specific effect corresponding to the set of blocked paths on G being the paths from X to the subset of its children who were the children of N on any given G_{ex} is identified under the NPSEM assumption for G . Their identifying formula is precisely our $f_{n=0,o=1}(v \setminus x)$ corresponding to this G_{ex} . In fact, our derivation implies that this path specific effect on G is identified by $f_{n=0,o=1}^{int}(v \setminus x)$ for this G_{ex} under the assumption that any of our four causal models associated with this G_{ex} holds, even without assuming that the NPSEM associated with the original graph G is true. Again under the NPSEM assumption for G , all 2^p effects $f_{n=0,o=1}^{int}(v \setminus x)$ as G_{ex} varies are identified, each by the formula $f_{n=0,o=1}(v \setminus x)$, specific to the graph G_{ex} .

A substantive scenario under which all 2^p effects $f_{n=0,o=1}^{int}(v \setminus x)$ are simultaneously identified by the graph G_{ex} specific-formulae $f_{n=0,o=1}(v \setminus x)$ is obtained by assuming an FFR-CISTG model for an expanded graph on which N and O are replaced by a set of parents X'_j , $j = 1, \dots, p$, one for each child of X , X is the only parent of each X'_j , each X'_j has a single child, and $X = X'_j$ with probability one in the actual data. We consider the 2^p interventions that set a subset of the X'_j to 1 and the remainder to 0. The relationship of this analysis to the analysis based on the graphs G_{ex} containing N and O mimics the relationship of the analysis based on Fig. 6(d) under the alternative scenario of the last subsection to the analyses based on Fig. 6(a) and Fig. 6(b). In these latter analyses, X had precisely three children: Z , L and Y .

Avin et al. (2005) also show that other path specific effects are identified under the NPSEM assumption for G . However their results imply that, whenever, for a given set of blocked paths, the path specific density of $V \setminus X$ is identified from data on V under an NPSEM associated with G , the identifying formula is equal to the g-formula density $f_{n=0,o=1}(v \setminus x)$ corresponding to one of the 2^p graphs G_{ex} . Avin et al. (2005) provide an algorithm that can be used to find the appropriate G_{ex} corresponding to a given set of blocked paths.

As discussed in Section 4.4 even the path-specific densities that are not identified under an NPSEM become identified under yet further untestable counterfactual independence

assumptions and/or rank preservation assumptions.

7 Conclusion

The results presented here, which are summarized in Table 1, appear to present a clear trade-off between the agnostic causal DAG, MCM and FFRCISTG model frameworks and that of the NPSEM.

In the NPSEM approach the PDE is identified, even though the result cannot be verified by a randomized experiment without making further assumptions. In contrast, the PDE is not identified under an agnostic causal DAG model or under an MCM / FFRCISTG model. Further, in the Appendix we show that the ETT can be identified under an MCM / FFRCISTG model even though the ETT cannot be verified by a randomized experiment without making further assumptions.

Our analysis of Pearl’s motivation for the PDE suggests that the above dichotomies may not be as stark as they may at first appear. We have shown that in certain cases where one is interested in a *prima facie* non-manipulable causal parameter then the very fact that it is of interest implies that there also exists an extended DAG in which the same parameter is manipulable and identifiable in all the causal frameworks.

Inevitably such cases will be interpreted differently by NPSEM ‘skeptics’ and ‘advocates’. Advocates may argue that if our conjecture holds then we can work with NPSEMs and have some reassurance that in important cases of scientific interest we will have the option to go back to an agnostic causal DAG. Conversely skeptics may conclude that if we are correct then this shows that it is advisable to avoid the NPSEM framework: agnostic causal DAGs are fully “testable” (with the usual caveats) and many non-manipulable NPSEM parameters that are of interest, but not identifiable within a non-NPSEM framework, can be identified in an augmented agnostic causal DAG.

Undoubtedly this debate is set to run and run . . .

Causal Model	Potential Outcome Indep. Ass.	Direct Effects								ETT	
		CDE		PDE		PSDE		\mathcal{X} =2		\mathcal{X} >2	
		D	I	D	I	D	I	D	I	D	I
Agnostic DAG	None	Y	Y	N	N	N	N	N	N	N	N
MCM	(5)	Y	Y	Y	N	Y	N	Y	Y	Y	N
FFRCISTG	(1)	Y	Y	Y	N	Y	N	Y	Y	Y	Y
NPSEM	(8)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 1: Relations between causal models and estimands associated with the DAG shown in Figure 1; column ‘D’ indicates if the contrast is defined in the model; ‘I’ whether it is identified.

A The Effect of Treatment on the Treated: a Non-Manipulable Parameter

The primary focus of this paper has been various contrasts assessing the direct effect of X on Y relative to an intermediate Z . In this appendix we discuss another non-manipulable parameter, the effect of treatment on the treated, in order to further clarify the differences among the agnostic, the MCM and the FFRCISTG models. For our purposes, we shall only require the simplest possible causal model based on the DAG $X \rightarrow Y$, obtained by marginalizing over Z in the graph in Figure 1. Let $Y(0)$ denote the counterfactual $Y(x)$ evaluated at $x = 0$. In a counterfactual causal model, the average effect of treatment on the treated is defined to be:

$$\text{ETT}(x) \equiv E[Y(x) - Y(0) \mid X = x] \equiv E_x^{\text{int}}[Y \mid X = x] - E_0^{\text{int}}[Y \mid X = x].$$

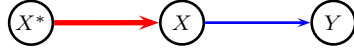


Figure 7: An extended DAG, with a treatment $X = X^*$ and response Y , leading to an interventional interpretation of the effect of treatment on the treated (Geneletti and Dawid, 2007; Robins et al., 2007).

A.1 Minimal Counterfactual Models (MCMs)

In an MCM associated with DAG $X \rightarrow Y$, $E[Y(x) | X = x] = E[Y | X = x]$, by the consistency assumption (iii) in §2.1. Thus

$$\text{ETT}(x) = E[Y | X = x] - E[Y(0) | X = x].$$

Hence the $\text{ETT}(x)$ is identified iff the second term on the right is identified. First note that

$$\text{ETT}(0) = E[Y | X = 0] - E[Y(0) | X = 0] = 0.$$

Now by consistency condition (iii) in §2.1 and the MCM assumption, Eq. (4), we have:

$$E[Y | X = 0] = E[Y(0) | X = 0] = E[Y(0)].$$

By the law of total probability

$$E[Y(0)] = \sum_x E[Y(0) | X = x] \Pr(X = x).$$

Hence it follows that

$$E[Y | X = 0] \Pr(X \neq 0) = \sum_{x:x \neq 0} E[Y(0) | X = x] \Pr(X = x). \quad (28)$$

In the special case where X is binary, so $|\mathcal{X}| = 2$, the right-hand side of Eq. (28) reduces to a single term and thus we have $E[Y(0) | X = 1] = E[Y | X = 0]$. It follows that for binary X , we have

$$\text{ETT}(1) = E[Y | X = 1] - E[Y | X = 0].$$

under the MCM (and hence any counterfactual causal model).

In contrast, if X is not binary, then the right-hand side of Eq. (28) contains more than one unknown so that $\text{ETT}(x)$ for $x \neq 0$ is not identified under the MCM.

However, under an FFRCISTG model, Eq. (1) implies that

$$E[Y(0) | X = x] = E[Y | X = 0],$$

so $\text{ETT}(x)$ is identified in this model, regardless of X 's sample space. The parameter $\text{ETT}(1) = E[Y(1) - Y(0) | X = 1]$ is not manipulable, relative to $\{X, Y\}$, even when X is binary, since, without further assumptions, we cannot experimentally observe $Y(0)$ in subjects with $X = 1$.

Note that even under the MCM with $|\mathcal{X}| > 2$, the non-manipulable (relative to $\{X, Y\}$) contrast $E[Y(0)|X \neq 0] - E[Y|X \neq 0]$, the effect of receiving $X = 0$ on those who did not receive $X = 0$, is identified, since $E[Y(0)|X \neq 0]$ is identified by the left hand side of (28).

We now turn to the agnostic causal model for the causal DAG $X \rightarrow Y$. Although $E_x^{int}[Y]$ is identified by the g-functional as $E[Y | X = x]$, nonetheless, as expected for a non-manipulable causal contrast, the effect of treatment on the treated is not formally defined within the agnostic causal model, without further assumptions, even for binary X . Of course, the g-functional (see Def. 5) does define a joint distribution $f_x(x^*, y)$ for (X, Y) under which X takes the value x with probability 1. However, in spite of apparent notational similarities, the conditional density, $f_x(y | x^*)$ expresses a different concept from that occurring in the definition of $E_x^{int}[Y | X = x^*] \equiv E[Y(x) | X = x^*]$ in the counterfactual theory. The former relates to the distribution over Y among those individuals who (after the intervention) have the value $X = x^*$, under an intervention which sets every unit's value to x and thus $f_x(y | x^*) = f(y | x)$ if $x^* = x$ and is undefined if $x^* \neq x$; the latter is based on the distribution of Y under an intervention fixing $X = x$ among those people who *would have had* the value $X = x^*$ had we not intervened.

The minimality of the MCM among all counterfactual models that both satisfy the consistency assumption (iii) in §2.1 and identify the intervention distributions $\{f_{PR}^{int}(z)\}$ can be seen as follows. For binary X , the above argument for identification of the non-manipulable contrast $\text{ETT}(1)$ under an MCM as the difference $E[Y | X = 1] - E[Y | X = 0]$ follows directly, via the laws of probability, from the consistency assumption (iii) in §2.1 and

the minimal independence assumption (5) required to identify the intervention distributions $\{f_{PR}^{int}(z)\}$. In contrast, the additional independence assumptions (8) used to identify the PDE under the NPSEM for the DAG in Figure 1 or the additional independence assumptions used to identify ETT(1) for non-binary X under a FFRCISTG model for the DAG $X \rightarrow Y$ are not needed to identify intervention distributions.

Of course, as we have shown, it may be the case that the PDE is identified as an intervention contrast in an extended causal DAG containing additional variables; but identification in this extended causal DAG requires additional assumptions beyond those in the original DAG and hence does not follow merely from the application of the laws of probability.

Similarly, the ETT(1) for the causal DAG $X \rightarrow Y$ can be re-interpreted as an intervention contrast in an extended causal DAG containing additional variables, regardless of the dimension of X 's state space. Specifically, Robins et al. (2007) showed the ETT(x) parameter is defined and identified via the extended agnostic causal DAG in Figure 7 that adds to the DAG $X \rightarrow Y$ a variable X^* that is equal to X with probability one under $f(v) = f(x^*, x, y)$. Then $E_x^{int}[Y|X^* = x^*]$ is identified by the g-formula as $E[Y|X = x]$, because X is the only parent of Y . Furthermore $E_x^{int}[Y|X^* = x^*]$ has an interpretation as the effect on the mean of Y of setting X to x on those observed to have $X = x^*$, because $X = X^*$ with probability one. Thus, though ETT(x) is not a manipulable parameter relative to the graph $X \rightarrow Y$, it is manipulable relative to the variables $\{X^*, X, Y\}$ in the DAG in Figure 7. In the extended graph ETT(x) is identified by the same function of the observed data as $E[Y(x) - Y(0) | X = x]$ in the original FFRCISTG model for non-binary X or in the original MCM or FFRCISTG model for binary X .

A.2 A model that is an MCM but not an FFRCISTG

In this section we describe a parametric counterfactual model for the effect of a ternary treatment X on a binary response Y that is an MCM associated with the graph in Figure 8, but is not an FFRCISTG. Let $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$, be a (vector valued) latent variable with 3 components such that in a given population $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_0, \alpha_1, \alpha_2)$, so that $\pi_0 + \pi_1 + \pi_2 = 1$ w.p. 1. The joint distribution of the factual and counterfactual data is determined by the unknown parameters $(\alpha_0, \alpha_1, \alpha_2)$. Specifically the treatment X is ternary with states 0, 1,

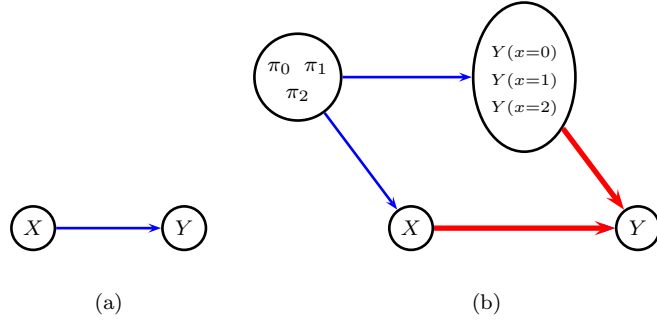


Figure 8: (a) A simple graph; (b) a graph describing confounding structure that leads to a counterfactual model that corresponds to the MCM, but not the FFRCISTG associated with the DAG (a); thicker red edges indicate deterministic relations.

2, and $P(X = k | \boldsymbol{\pi}) = \pi_k$, equivalently

$$X | \boldsymbol{\pi} \sim \text{Multinomial}(1, \boldsymbol{\pi}).$$

Now suppose that the response Y is binary and that the counterfactual outcomes $Y(x)$ are as follows:

$$Y(x = 0) | \boldsymbol{\pi} \sim \text{Bernoulli}(\pi_1 / (\pi_1 + \pi_2)),$$

$$Y(x = 1) | \boldsymbol{\pi} \sim \text{Bernoulli}(\pi_2 / (\pi_2 + \pi_3)),$$

$$Y(x = 2) | \boldsymbol{\pi} \sim \text{Bernoulli}(\pi_0 / (\pi_0 + \pi_1)).$$

Thus in this example, conditional on $\boldsymbol{\pi}$ the potential outcome $Y(x = k)$ ‘happens’ to be a realization of a Bernoulli random variable with probability of success equal to the probability of receiving treatment $X = k + 1 \pmod 3$ given that treatment X is not k . In what follows we will use $[\cdot]$ to indicate that an expression is evaluated mod 3. Now since (π_0, π_1, π_2) follows a Dirichlet distribution, it follows that:

$$\pi_{[i+1]} / (\pi_{[i+1]} + \pi_{[i+2]}) \perp\!\!\!\perp \pi_i$$

for $i = 0, 1, 2$. Hence, in this example, for $i = 0, 1, 2$ we have $Y(x = i) \perp\!\!\!\perp \pi_i$. Further, $I(X = i) \perp\!\!\!\perp Y(x = i) | \pi_i$, hence the model obeys the MCM independence restriction (5):

$$Y(x = i) \perp\!\!\!\perp I(X = i) \text{ for all } i,$$

but not the FFRCISTG independence restriction (1), since:

$$Y(x = i) \not\perp\!\!\!\perp I(X = j) \text{ for } i \neq j.$$

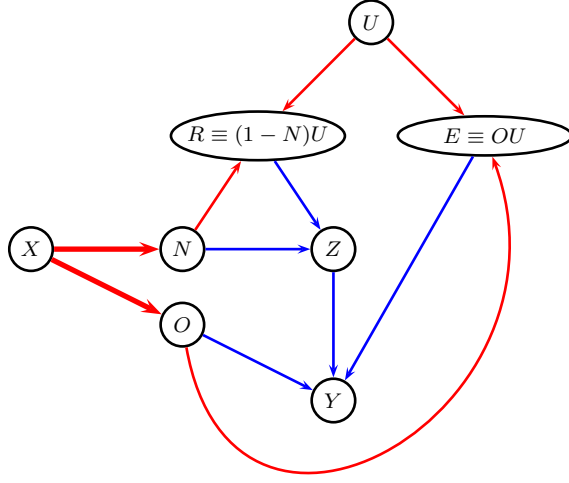


Figure 9: An example leading to the FFRCISTG associated with the DAG in Figure 1 but not an NPSEM.

We note that we have:

$$P(X = i) = E(\pi_i) = \alpha_i / (\alpha_0 + \alpha_1 + \alpha_2), \quad (29)$$

$$\boldsymbol{\pi} \mid X = i \sim \text{Dirichlet}(\alpha_i + 1, \alpha_{[i+1]}, \alpha_{[i+2]}), \quad (30)$$

$$Y(x = i) \mid X = i \sim \text{Bernoulli}(\alpha_{[i+1]} / (\alpha_{[i+1]} + \alpha_{[i+2]})), \quad (31)$$

$$Y \mid X = i \sim \text{Bernoulli}(\alpha_{[i+1]} / (\alpha_{[i+1]} + \alpha_{[i+2]})). \quad (32)$$

Equation (30) follows from standard Bayesian updating (since the Dirichlet distribution is conjugate to the multinomial). It follows that the vector of parameters $(\alpha_0, \alpha_1, \alpha_2)$ is identified only up to a scale factor, since the likelihood for the observed variables $p(x, y \mid \boldsymbol{\alpha}) = p(x, y \mid \lambda \boldsymbol{\alpha})$ for any $\lambda > 0$, by Eqs. (29) and (32). We note that since $E(Y(x)) = E(Y \mid X = x)$, $\text{ACE}_{X \rightarrow Y}(x) \equiv E(Y(x)) - E(Y(0)) = E(Y \mid X = x) - E(Y \mid X = 0)$, and thus is identified. However since

$$Y(x = 0) \mid X = 1 \sim \text{Bernoulli}((\alpha_1 + 1) / (\alpha_1 + 1 + \alpha_2)),$$

$$Y(x = 0) \mid X = 2 \sim \text{Bernoulli}(\alpha_1 / (\alpha_1 + \alpha_2 + 1)),$$

and the probability of success in these distributions is not invariant under rescaling of the vector $\boldsymbol{\alpha}$, we conclude that these distributions are not identified from data on $p(x, y)$. Consequently $\text{ETT}(x^*) \equiv E[Y(x = x^*) - Y(x = 0) \mid X = x^*]$ is not identified under our parametric model.

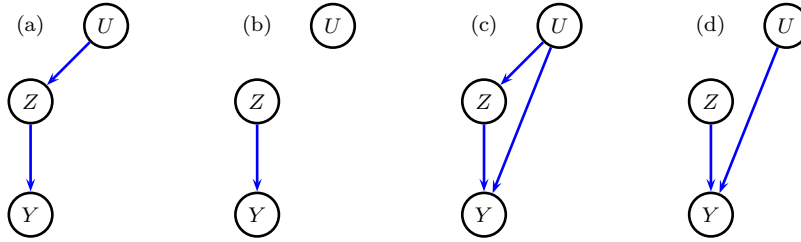


Figure 10: An example leading to the FFRCISTG associated with the DAG in Figure 1 holding but not the NPSEM: Causal subgraphs on U, Z, Y implied by the graph in Figure 9 when we intervene and set (a) $N = 0, O = 0$; (b) $N = 1, O = 0$; (c) $N = 0, O = 1$; (d) $N = 1, O = 1$.

Appendix B: A Data Generating Process Leading to an FFRCISTG that is not an NPSEM

Robins (2003) stated that it is hard to construct realistic (as opposed to mathematical) scenarios in which one would accept that the FFRCISTG model associated with Figure 1 held, but not the NPSEM, and thus that controlled direct effects are identified but pure direct effects are not. In this appendix we describe such a scenario. We leave it to the reader to judge its realism.

Suppose a substance U that is endogenously produced by the body could both (i) decrease blood pressure by reversibly binding to a membrane receptor on the blood pressure control cells in the carotid artery of the neck and (ii) directly increase atherosclerosis and thus MI by stimulating the endothelial cells of the coronary arteries of the heart via an interaction with a particular protein and (iii) the aforementioned protein is expressed in endothelial cells of the coronary arteries only when induced by the chemicals in tobacco smoke other than nicotine, e.g. tar. Further, suppose one mechanism by which nicotine increased blood pressure Z was by irreversibly binding to the membrane receptor for U on the blood pressure control cells in the carotid, the dose of nicotine in a smoker being sufficient to bind every available receptor. Then, under the assumption that there do not exist further unmeasured confounders for the effect of hypertension on MI, this scenario implies that it is reasonable

to assume that any of the four causal models associated with the expanded DAG in Figure 9 is true. Here R measures the degree of binding of protein U to the membrane receptor in blood pressure control cells. Thus R is zero in smokers of cigarettes containing nicotine. E measures the degree of stimulation of the endothelial cells of the carotid artery by U . Thus E is zero except in smokers (regardless of whether the cigarette contains nicotine).

Before considering whether the NPSEM associated with Figure 1 holds, let us first study the expanded DAG of Figure 9. An application of the g-formula to the DAG in Figure 9 shows that the effect of not smoking $E_{n=0,o=0}^{int}[Y] = E_{x=0}^{int}[Y]$ and the effect of smoking $E_{n=1,o=1}^{int}[Y] = E_{x=1}^{int}[Y]$ are identified by $E[Y | X = 0]$ and $E[Y | X = 1]$ under all four causal models associated with Figure 9. However, the effect $E_{n=0,o=1}^{int}[Y]$ of smoking nicotine-free cigarettes is not identified. Specifically

$$\begin{aligned}
& E_{n=0,o=1}^{int}[Y] \\
&= \sum_{z,u} E[Y | O = 1, U = u, Z = z] f(z | U = u, N = 0) f(u) \\
&= \sum_{z,u} E[Y | N = 1, O = 1, U = u, Z = z] f(z | U = u, N = 0, O = 0) f(u) \\
&= \sum_{z,u} E[Y | X = 1, U = u, Z = z] f(z | U = u, X = 0) f(u)
\end{aligned}$$

where the first equality used the fact that E is a deterministic function of U and O and that R is a deterministic function of N and U . The second equality used d-separation and the third determinism. Thus $E_{n=0,o=1}^{int}[Y]$ is not a function of the density of the observed data on (X, Z, Y) , because u occurs both in the term $E[Y | X = 1, U = u, Z = z]$ where we have conditioned on $X = 1$ and the term $f(z | U = u, X = 0)$ where we have conditioned on $X = 0$. As a consequence, we do not obtain a function of the density of the observed data when we marginalize over u .

Since, under all three counterfactual models associated with the extended DAG of Figure 9, $E_{n=0,o=1}^{int}[Y]$ is equal to the parameter $E[Y(x = 1, Z(x = 0))]$ of Figure 1, we conclude that $E[Y(x = 1, Z(x = 0))]$ and thus the PDE is not identified. Hence the induced counterfactual model for the DAG in Figure 1 cannot be an NPSEM (as that would imply that the PDE would be identified).

Furthermore, $E_{n=0,o=1}^{int}[Y]$ is a manipulable parameter with respect to the DAG in Figure

3, since this DAG is obtained from marginalizing over U in the graph in Figure 9. However, as we showed above, $E_{n=0,o=1}^{int}[Y]$ is not identified from the law of the factu- als X, Y, Z, N, O , which are the variables in Figure 3. From this we conclude that none of the four causal models associated with the graph in Figure 3 can be true. Note that *prima facie* one might have thought that if the agnostic causal DAG in Figure 1 is true, then this would always imply that the agnostic causal DAG in Figure 3 is also true. This example demonstrates that such a conclusion is fallacious. Similar remarks apply to the FFRCISTG models.

Furthermore, for $z = 0, 1$, by applying the g-formula to the graph in Fig. 9, we obtain that the joint effect of smoking and z , $E_{n=1,o=1,z}^{int}[Y]$, and the joint effect of not smoking and z , $E_{n=0,o=0,z}^{int}[Y]$, are identified by $E[Y|X = 1, Z = z]$ and $E[Y|X = 0, Z = z]$, respectively, under all four causal models for Figure 9. Since $E_{n=0,o=0,z}^{int}[Y]$ and $E_{n=1,o=1,z}^{int}[Y]$ are equal to the parameters $E_{x=0,z}^{int}[Y]$ and $E_{x=1,z}^{int}[Y]$ under all four associated causal models associated with the graph in Figure 1, we conclude $CDE(z)$ is also identified under all four causal models associated with Figure 1.

The results obtained in the last two paragraphs are consistent with the FFRCISTG associated with the graph in Figure 1 holding, but not the NPSEM. Below we prove such is the case.

Before doing so we provide a simpler and more intuitive way to understand the above results by displaying in Figure 10 the subgraphs of Figure 9 corresponding to U, Z, Y when the variables N and O are set to each of their 4 possible joint values. We see that only when we set $N = 0, O = 1$ is U a common cause of both Z and Y (as setting $N = 0, O = 1$ makes $R = E = U$). Thus we have

$$\begin{aligned} E_{n=0,o=0,z}^{int}[Y] &= E_{n=0,o=0}^{int}[Y|Z = z] \\ &= E[Y|O = 0, N = 0, Z = z] = E[Y|X = 0, Z = z], \text{ and} \\ E_{n=1,o=1,z}^{int}[Y] &= E_{n=1,o=1}^{int}[Y|Z = z] \\ &= E[Y|O = 1, N = 1, Z = z] = E[Y|X = 1, Z = z] \end{aligned}$$

as O and N are unconfounded and Z is unconfounded when either we set $O = 1, N = 1$ or we set $O = 0, N = 0$. However $E_{n=0,o=1,z}^{int}[Y] \neq E_{n=0,o=1}^{int}[Y|z] = E[Y|N = 0, O = 1, Z = z]$ as the effect of Z on Y is confounded when we set $N = 0, O = 1$. It is because $E_{n=0,o=1,z}^{int}[Y] \neq$

$E_{n=0,o=1}^{int} [Y|z]$ that $E_{n=0,o=1}^{int} [Y]$ is not identified. If, contrary to Figure 9, there was no confounding between Y and Z when N is set to 0 and O is set to 1 then we would have $E_{n=0,o=1,z}^{int} [Y] = E_{n=0,o=1}^{int} [Y|z]$. It would then follow that

$$\begin{aligned}
E_{n=0,o=1}^{int} [Y] &= \sum_z E_{n=0,o=1}^{int} [Y|z] f_{n=0,o=1}^{int} [z] \\
&= \sum_z E_{n=0,o=1,z}^{int} [Y] f_{n=0,o=1}^{int} [z] \\
&= \sum_z E_{n=1,o=1,z}^{int} [Y] f_{n=0,o=0}^{int} [z] \\
&= \sum_z E [Y|X = 1, Z = z] f [z|X = 0],
\end{aligned}$$

where the third equality is from the fact that we suppose N has no direct effect on Y not through Z and O has no effect on Z .

We conclude by showing that the MCM and FFRCISTG models associated with Figure 1 are true but the NPSEM is not, if any of the three counterfactual models associated with Figure 9 are true. Specifically, the DAG in Figure 11 represents the DAG of Figure 1 with the counterfactuals for $Z(x)$ and $Y(x, z)$, the variable U of Figure 9, and common causes U_1 and U_2 of the $Z(x)$ and the $Y(x, z)$ added to the graph. Note that U being a common cause of Z and Y in Figures 9 and 10 only when we set $N = 0$ and $O = 1$ implies that U is only a common cause of $Z(0)$, $Y(1, 0)$ and $Y(1, 1)$ in Figure 11. One can check using d-separation that the counterfactual independencies in Figure 11 satisfy those required of an MCM or FFRCISTG model, but not those of a NPSEM, as $Z(0)$ and $Y(1, z)$ are dependent. However, Figure 11 contains more independencies than are required for the FFRCISTG condition (1) applied to the DAG in Figure 1. In particular, in Figure 11, $Z(1)$ and $Y(0, z)$ are independent which implies that $E [Y(0, Z(1))]$ is identified by $\sum_z E [Y|X = 0, Z = z] f(z|X = 1)$ and thus the the so-called total direct effect $E [Y(1, Z(1))] - E [Y(0, Z(1))]$ is also identified. Finally, we note that we could easily modify our example to eliminate the independence of $Z(1)$ and $Y(0, z)$.

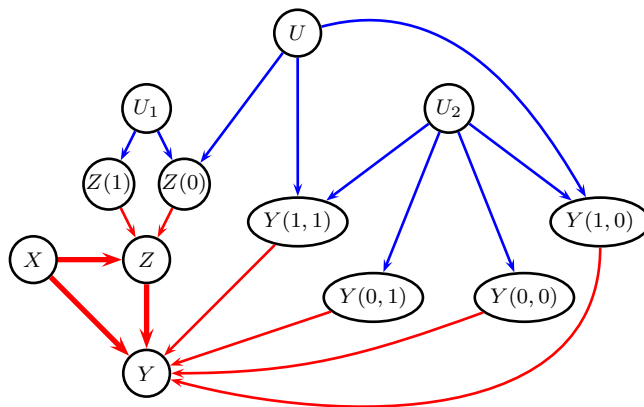


Figure 11: An example leading to an FFRCISTG corresponding to the DAG in Figure 1 but not an NPSEM: potential outcome perspective. Red edges indicate that the variable is a deterministic function of its parents; blue indicate a stochastic relation. Counterfactuals for Y are indexed $Y(x, z)$. U , U_1 and U_2 indicate hidden confounders.

Appendix C: Bounds on the PDE under an FFRCISTG Model

In this Appendix we derive bounds on the PDE

$$\text{PDE} = E[Y(x = 1, Z(x = 0))] - E[Y | X = 0]$$

under the assumption that the MCM or FFRCISTG model corresponding to the graph in Figure 1 holds, and all variables are binary. Note

$$\begin{aligned} & E[Y(x = 1, Z(x = 0))] \\ &= E[Y(x = 1, z = 0) | Z(x = 0) = 0]p(Z = 0 | X = 0) \\ &\quad + E[Y(x = 1, z = 1) | Z(x = 0) = 1]p(Z = 1 | X = 0). \end{aligned}$$

The two quantities $E[Y(x = 1, z = 0) | Z(x = 0) = 0]$ and $E[Y(x = 1, z = 1) | Z(x = 0) = 1]$ are constrained by the law for the observed data via:

$$\begin{aligned}
E[Y | X=1, Z=0] &= E[Y(x = 1, z = 0)] \\
&= E[Y(x = 1, z = 0) | Z(x = 0) = 0]p(Z(x = 0) = 0) \\
&\quad + E[Y(x = 1, z = 0) | Z(x = 0) = 1]p(Z(x = 0) = 1) \\
&= E[Y(x = 1, z = 0) | Z(x = 0) = 0]p(Z = 0 | X = 0) \\
&\quad + E[Y(x = 1, z = 0) | Z(x = 0) = 1]p(Z = 1 | X = 0),
\end{aligned}$$

$$\begin{aligned}
E[Y | X=1, Z=1] &= E[Y(x = 1, z = 1)] \\
&= E[Y(x = 1, z = 1) | Z(x = 0) = 0]p(Z(x = 0) = 0) \\
&\quad + E[Y(x = 1, z = 1) | Z(x = 0) = 1]p(Z(x = 0) = 1) \\
&= E[Y(x = 1, z = 1) | Z(x = 0) = 0]p(Z = 0 | X = 0) \\
&\quad + E[Y(x = 1, z = 1) | Z(x = 0) = 1]p(Z = 1 | X = 0).
\end{aligned}$$

It then follows from the analysis in §2.2 in Richardson and Robins (2010) that the set of possible values for the pair

$$(\alpha_0, \alpha_1) \equiv (E[Y(x=1, z=0) | Z(x=0) = 0], E[Y(x=1, z=1) | Z(x=0) = 1])$$

compatible with the observed joint distribution $p(z, y | x)$ is given by:

$$(\alpha_0, \alpha_1) \in [l_0, u_0] \times [l_1, u_1]$$

where,

$$l_0 = \max\{0, 1 + (E[Y | X = 1, Z = 0] - 1)/p(Z = 0 | X = 0)\},$$

$$u_0 = \min\{E[Y | X = 1, Z = 0]/p(Z = 0 | X = 0), 1\},$$

$$l_1 = \max\{0, 1 + (E[Y | X = 1, Z = 1] - 1)/p(Z = 1 | X = 0)\},$$

$$u_1 = \min\{E[Y | X = 1, Z = 1]/p(Z = 1 | X = 0), 1\}.$$

Hence we have the following upper and lower bounds on the PDE:

$$\begin{aligned}
& \max\{0, p(Z=0|X=0) + E[Y|X=1, Z=0] - 1\} + \\
& \max\{0, p(Z=1|X=0) + E[Y|X=1, Z=1] - 1\} - E[Y | X = 0] \\
& \leq \text{PDE} \leq \\
& \min\{p(Z=0|X=0), E[Y|X=1, Z=0]\} + \\
& \min\{p(Z=1|X=0), E[Y|X=1, Z=1]\} - E[Y | X = 0].
\end{aligned}$$

Kaufman et al. (2009) obtain bounds on the PDE under assumption (2) but while allowing for confounding between Z and Y , i.e. not assuming that (3) holds, as we do. As we would expect the bounds that we obtain are strictly contained those obtained by Kaufman et al.; see Table 2, row {5'} in Kaufman et al. (2009). Note, when $p(Z = z | X = 0) = 1$, $\text{PDE} = \text{CDE}(z) = E[Y|X = 1, Z = z] - E[Y|X = 0, Z = z]$ and thus our upper and lower bounds on the PDE coincide. In contrast, Kaufman et al.'s upper and lower bounds on the PDE do not coincide when $p(Z = z | X = 0) = 1$ and thus $\text{PDE} = \text{CDE}(z)$, as the $\text{CDE}(z)$ is not identified under their assumptions.

Appendix D: Interventions Restricted to a Subset: The FRCISTG Model

To describe the FRCISTG model for $V = (V_1, \dots, V_M)$, we suppose that each $V_m = (L_m, A_m)$ is actually a composite of variables L_m and A_m , one of which can be the empty set. The causal effects of intervening on the any of the L_m variables is not defined. However, we assume that for any subset R of $A = \bar{A}_M = (A_1, \dots, A_M)$, the counterfactuals $V_m(r)$ are well-defined for any $r \in \mathcal{R}$.

Specifically, we assume that the one step ahead counterfactuals $V_m(\bar{a}_{m-1}) = (L_m(\bar{a}_{m-1}), A_m(\bar{a}_{m-1}))$ exist for any setting of $\bar{a}_{m-1} \in \bar{\mathcal{A}}_{m-1}$. Note it is implicit in this definition that L_k precedes A_k for all k . Next we make the consistency assumption that the factual variables V_m and counterfactual variables $V_m(r)$ are obtained recursively from the $V_m(\bar{a}_{m-1})$. We do not provide a graphical characterization of parents. Rather, we say that the parents Pa_m of V_m

consist of the smallest subset of \bar{A}_{m-1} such that, for all $\bar{a}_{m-1} \in \bar{A}_{m-1}$, $V_m(\bar{a}_{m-1}) = V_m(pa_m)$ where pa_m is the sub-vector of \bar{a}_{m-1} corresponding to Pa_m . One can then view the parents Pa_m of V_m as the direct causes of V_m relative to the variables prior to V_m on which we can perform interventions. Finally, an FRCISTG model imposes the following independencies:

$$\{V_{m+1}(\bar{a}_m), \dots, V_M(\bar{a}_{M-1})\} \perp\!\!\!\perp A_m(\bar{a}_{m-1}) \mid \bar{L}_m = \bar{l}_m, \bar{A}_{m-1} = \bar{a}_{m-1}, \quad (33)$$

for all $m, \bar{a}_{M-1}, \bar{l}_m$.

Note that (33) can also be written

$$\{V_{m+1}(\bar{a}_m), \dots, V_M(\bar{a}_{M-1})\} \perp\!\!\!\perp A_m(\bar{a}_{m-1}) \mid \bar{L}_m(\bar{a}_{m-1}) = \bar{l}_m, \bar{A}_{m-1} = \bar{a}_{m-1},$$

for all $m, \bar{a}_{M-1}, \bar{l}_m$,

where $\bar{L}_m(\bar{a}_{m-1}) = (L_m(\bar{a}_{m-1}), L_{m-1}(\bar{a}_{m-2}), \dots, L_1)$.

In the absence of inter-unit interference and non-compliance, data from a sequentially randomized experiment in which at each time m , the treatment A_m is randomly assigned, with the assignment probability at m possibly depending on the past $(\bar{L}_m, \bar{A}_{m-1})$, will follow an FRCISTG model. See Robins (1986) for further discussion.

The analogous *minimal causal model (MCM) with interventions restricted to a subset* is defined by replacing $A_m(\bar{a}_{m-1})$ by $I\{A_m(\bar{a}_{m-1}) = a_m\}$ in condition (33).

It follows from Robins (1986) that our Extended Lemma 6 continues to hold when we substitute either ‘FRCISTG model’ or ‘MCM with restricted interventions’, for ‘MCM’ in the statement of the Lemma, provided we take $R \subseteq A$.

Likewise we may define an *agnostic causal model with restricted interventions* to be the causal model that simply assumes that the interventional density of $Z \subset V$, denoted by $f_{p_R}^{int}(z)$, under treatment regime p_R for any $R \subseteq A$, is given by the g-functional density $f_{p_R}(z)$, whenever $f_{p_R}(z)$ is a well-defined function of $f(v)$.

In Theorem 1 we proved that the set of defining conditional independencies in Eq. (1) of an FFRICISTG model can be re-expressed as a set of unconditional independencies between counterfactuals. An analogous result does not hold for an FRCISTG. However the following theorem shows that we can remove past treatment history from the conditioning set in the defining conditional independencies of an FRCISTG model, provided that we continue to condition on the counterfactuals $\bar{L}_m(\bar{a}_{m-1})$.

Theorem 8 An FRCISTG model for $V = (V_1, \dots, V_M)$, $V_m = (L_m, A_m)$ implies that for all m , \bar{a}_{M-1}, \bar{l}_m ,

$$\{V_{m+1}(\bar{a}_m), \dots, V_M(\bar{a}_{M-1})\} \perp\!\!\!\perp A_m(\bar{a}_{m-1}) \mid \bar{L}_m(\bar{a}_{m-1}) = \bar{l}_m.$$

Note the theorem would not be true had we substituted the factual \bar{L}_m for $\bar{L}_m(\bar{a}_{m-1})$.

References

- Avin, C., I. Shpitser, and J. Pearl (2005). Identifiability of path-specific effects. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pp. 357–363.
- Dawid, A. (2000a). Causal inference without counterfactuals (with discussion). *J. Amer. Statist. Assoc.* 95, 407–448.
- Dawid, A. P. (2000b). Causal inference without counterfactuals (C/R: P424-448). *Journal of the American Statistical Association* 95(450), 407–424.
- Didelez, V., A. Dawid, and S. Geneletti (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, pp. 138–146. AUAI Press.
- Frangakis, C. and D. Rubin (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86(2), 365–379.
- Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics* 58(1), 21–29.
- Geneletti, S. and A. P. Dawid (2007). Defining and identifying the effect of treatment on the treated. Technical Report 3, Department of Epidemiology and Public Health, Imperial College London.
- Gill, R. D. and J. M. Robins (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* 29(6), 1785–1811.

- Hafeman, D. and T. VanderWeele (2009). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* $\varnothing(?)$, $??-??$
- Heckerman, D. and R. D. Shachter (1995). A definition and graphical representation for causality. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, August 18-20, 1995, Montreal, Quebec, Canada*, pp. 262–273.
- Imai, K., L. Keele, and T. Yamamoto (2009). Identification, inference, and sensitivity analysis for causal mediation effects. Technical report, Department of Politics, Princeton University.
- Kaufman, S., J. S. Kaufman, and R. F. MacLehose (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*. doi: 10.1016/j.jspi.2009.03.024.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.
- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, San Francisco, CA, pp. 411–42. Morgan Kaufmann.
- Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics* 6(2). DOI: 10.2202/1557-4679.1203.
- Petersen, M., S. Sinisi, and M. van der Laan (2006). Estimation of direct causal effects. *Epidemiology* 17(17), 276–284.
- Richardson, T. S. and J. M. Robins (2010). Analysis of the binary instrumental variable model. In R. Dechter, H. Geffner, and J. Halpern (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, Chapter 25, pp. 415–444. London: College Publications.

- Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling* 7, 1393–1512.
- Robins, J. (1987). Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect”. *Computers and Mathematics with Applications* 14, 923–945.
- Robins, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 70–81. Oxford, UK: Oxford University Press.
- Robins, J., A. Rotnitzky, and S. Vansteelandt (2007). Discussion of “Principal stratification designs to estimate input data missing due to death” by Frangakis, C.E., Rubin D.B., An, M., MacKenzie, E. *Biometrics* 63(3), 650–653.
- Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3, 143–155.
- Robins, J. M. and S. Greenland (2000). Comment on “Causal inference without counterfactuals”. *Journal of the American Statistical Association* 95(450), 431–435.
- Robins, J. M., T. S. Richardson, and P. Spirtes (2009). Identification and inference for direct effects. Technical Report 563, Department of Statistics, University of Washington.
- Robins, J. M., T. J. VanderWeele, and T. S. Richardson (2007). Discussion of “Causal effects in the presence of non compliance a latent variable interpretation” by Forcina, A. *Metron* LXIV(3), 288–298.
- Rothman, K. J. (1976). Causes. *Am. J. Epidemiol.* 104, 587–592.
- Rubin, D. B. (1998). More powerful randomization-based p -values in double-blind trials with non-compliance. *Statistics in Medicine* 17, 371–385.
- Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scand. J. Statist.* 31(2), 161–170.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*. Number 81 in Lecture Notes in Statistics. Springer-Verlag.