Data, Design, and Background Knowledge in Etiologic Inference

James M. Robins

I use two examples to demonstrate that an appropriate etiologic analysis of an epidemiologic study depends as much on study design and background subject-matter knowledge as on the data. The demonstration is facilitated by the use of causal graphs. (Epidemiology 2001;11:313–320)

Key Words: inference, etiology, study design, data collection, data analysis, epidemiologic methods

Greenland *et al*[1] discussed the use of causal graphs in epidemiologic research. A limitation of that paper was that it was lacking concrete examples designed to help the reader see how to take one's knowledge of study design, temporal ordering, basic biology, and epidemiologic principles to construct an appropriate causal graph. Here I present two epidemiologic thought experiments that make the point that the choice of an appropriate etiologic analysis depends as much on the design of the study and background subject-matter knowledge as on the data.

Specifically, in the first, I provide a single hypothetical dataset and three differing study designs, each of which plausibly could have given rise to the data. I show that the appropriate etiologic analysis differs with the design. In the second, I revisit a well-known epidemiologic controversy from the late 1970s. Horowitz and Feinstein[2] proposed that the strong association between postmenopausal estrogens and endometrial cancer seen in many epidemiologic studies might be wholly attributable to diagnostic bias. Others disagreed.[3–5] Part of the discussion centered on the issue of whether it was appropriate to stratify on vaginal bleeding, the purported cause of the diagnostic bias in the analysis. The goal here is to show, using causal graphs, that the answer depends on underlying assumptions about the relevant biological mechanisms.

From the Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115.

Address correspondence to: James M. Robins, Harvard School of Public Health, Boston, MA 02115.

Submitted June 15, 2000; final version accepted October 26, 2000.

1. Thought Experiment 1

Consider the data given in Table 1. E is a correctly classified exposure of interest whose net causal effect on a disease outcome D I would like to ascertain. E^* is a misclassified version of E. We are interested in the effect of E on D. Data on E, E^*, and D are available on all study subjects. Sampling variability can be ignored. I will now describe the designs of three different studies. For each study, the data are the same. Only the designs are different. I wish to answer the following questions for each of the studies: Can one say whether exposure has an adverse, protective, or no causal effect on the outcome? What association measure is most likely to have a causal interpretation?

As a guide, I present some candidate association measures. In Table 2, I calculate the exposure-disease odds ratio $OR_{ED} = 1.73$. I can also calculate the conditional ED odds ratio within strata of E^*, that is, $OR_{ED|E^*=1} = OR_{ED|E^*=0} = 3$. Similarly, I calculate that $OR_{E^*D} = 0.5$ and $OR_{E^*D|E=1} = OR_{E^*D|E=0} = 0.3$. I will report all associations on an odds ratio scale. This choice is dictated by the fact that in the case-control study described below, the only estimable population association measures are odds ratios.

Case-Control Study

Suppose the data arose from a case-control study of the effect of a particular nonsteroidal anti-inflammatory drug (E) on a congenital defect (D) that arises in the second trimester. Cases ($D = 1$) are infants with the congenital defect. Controls ($D = 0$) are infants without the defect. The control sampling fraction is unknown. The data E^* were obtained 1 month postpartum by maternal self-report. The data E were obtained from comprehensive accurate medical records of first trimester medications. All relevant preconception confounders and other drug exposures were controlled by stratification. The data in Table 1 are taken from a particular

TABLE 1. Data from a Hypothetical Study

	$D = 1$			$D = 0$	
	$E^* = 1$	$E^* = 0$		$E^* = 1$	$E^* = 0$
$E = 1$	180	200	$E = 1$	600	200
$E = 0$	20	200	$E = 0$	200	600

stratum. Note that misclassification is differential, given that $OR_{E^*D \mid E=1} = OR_{E^*D \mid E=0} = 0.3 \neq 1$.

Prospective Cohort Study

Suppose the data were obtained from a follow-up study of total mortality (D) in a cohort of short-term healthy 25-year-old uranium miners, all of whom only worked underground in 1967 for 6 months. The follow-up is complete through 1997. Suppose, for simplicity, there is a threshold pulmonary dose below which exposure to radon is known to have no effect on mortality. Let $E = 1$ ($E = 0$) denote above-threshold (below-threshold) exposure to radon as measured by lung dosimetry. Each miner was also assigned an estimated radon exposure E^* on the basis of the air level of radon in his mine. Let $E^* = 1$ ($E^* = 0$) denote an estimate above (below) threshold radon exposure. The assignment of miners to particular mines was unrelated to lifestyle, demographic, or medical risk factors. A subject's actual exposure E depends both on the level of radon in the mine and on the demands of the subject's job, such as the required amount of physical exertion and thus minute ventilation. Finally, it is known that 6 months of physical exertion at age 25 has no independent effect on later mortality.

Randomized Clinical Trial

Suppose the data were obtained from a randomized follow-up study of the effect of low-fat diet on death (D) over a 15-year follow-up period. Study subjects were randomly assigned to either a low-fat diet, educational, and motivational intervention arm ($E^* = 1$) or to a standard care arm ($E^* = 0$). Investigators were able to obtain accurate measures of the actual diet followed by the study subjects: $E = 1$ if a study subject followed a low-fat diet, and $E = 0$ otherwise. Assume E^* has no direct effect on death (D) except through its effect on actual fat consumption E.

Causal Contrasts

To determine which association measure is most likely causal, I need a formal definition of causal effects. Causal effects are best expressed in terms of counterfactual variables. Let the variable $D(1)$ denote a subject's outcome if exposed and $D(0)$ denote a subject's outcome if unexposed. For a given subject, the causal effect of treatment, measured on a difference scale, is $D(1) - D(0)$. If a subject is exposed ($E = 1$), the subject's observed outcome D equals $D(1)$, and $D(0)$ is unobserved. If $E = 0$, D equals $D(0)$, and $D(1)$ is unobserved. Let $pr[D(1) = 1]$ and $pr[D(0) = 1]$, respectively, be the probability that $D(1)$ is equal to 1 and $D(0)$ is equal to 1, where probabilities refer to proportions in a large, possibly hypothetical, source population. Then, the exposure-disease causal odds ratios is $OR_{causal,ED} = \{pr[D(1) = 1]/pr[D(1) = 0]\}/\{pr[D(0) = 1]/pr[D(0) = 0]\} = pr[D(1) = 1]pr[D(0) = 0]/\{pr[D(1) = 0]pr[D(0) = 1]\}$. For any variable Z, the exposure-disease causal odds ratio among the subset of subjects with Z being z is $OR_{causal,ED \mid Z=z} = pr[D(1) = 1 \mid Z = z]pr[D(0) = 0 \mid Z = z]/\{pr[D(1) = 0 \mid Z = z]pr[D(0) = 1 \mid Z = z]\}$

.

Answers

In this subsection, we provide the appropriate answers. The justification for these answers is given after I have reviewed causal graphs below. In the case-control study, exposure is likely harmful and the best parameter choice is the crude odds ratio $OR_{DE} = 1.73$. The other measures are biased. In particular, the conditional odds ratio $OR_{ED \mid E^*} = 3$ is biased in the sense that it fails to equal the causal effect $OR_{causal,ED \mid E^*}$ of exposure on disease among subjects within a particular stratum of E^*.

In the prospective cohort study, exposure is likely beneficial, and the best parameter choice is the conditional odds ratio $OR_{DE \mid E^*} = 3$. In the randomized trial, exposure is likely beneficial, and the best parameter choice may be the crude E^*D association $OR_{E^*D} = 0.5$, although it is likely that this association underestimates the true benefit of exposure. In this case, both the crude association $OR_{ED} = 1.73$, and the conditional association $OR_{ED \mid E^*} = 3$ are biased estimates of the causal effect of E on D. These answers clearly show that the appropriate statistical analysis depends on the design.

Causal Graphs

To justify the answers, we review causal directed acyclic graphs (DAGs) as discussed by Pearl and Verma,[6] Spirtes *et al*,[7] Pearl,[8] Pearl and Robins,[9] and Greenland *et al*.[1]

A causal graph is a directed acyclic graph (DAG) in which the vertices (nodes) of the graph represent variables; the directed edges (arrows) represent direct causal relations between variables; and there are no directed cycles, because no variable can cause itself (Figure 1). For a DAG to be causal, the variables represented on the graph must include the measured variables and additional unmeasured variables, such that if any two variables on the graph have a cause in common, that common cause is itself included as a variable on the graph. For example, in DAG 1, E and D are the measured variables. U represents all unmeasured common causes of E and D.

TABLE 2. Crude Data from a Hypothetical Study

	$E = 1$	$E = 0$
$D = 1$	380	220
$D = 0$	800	800
	$OR = 1.73$	

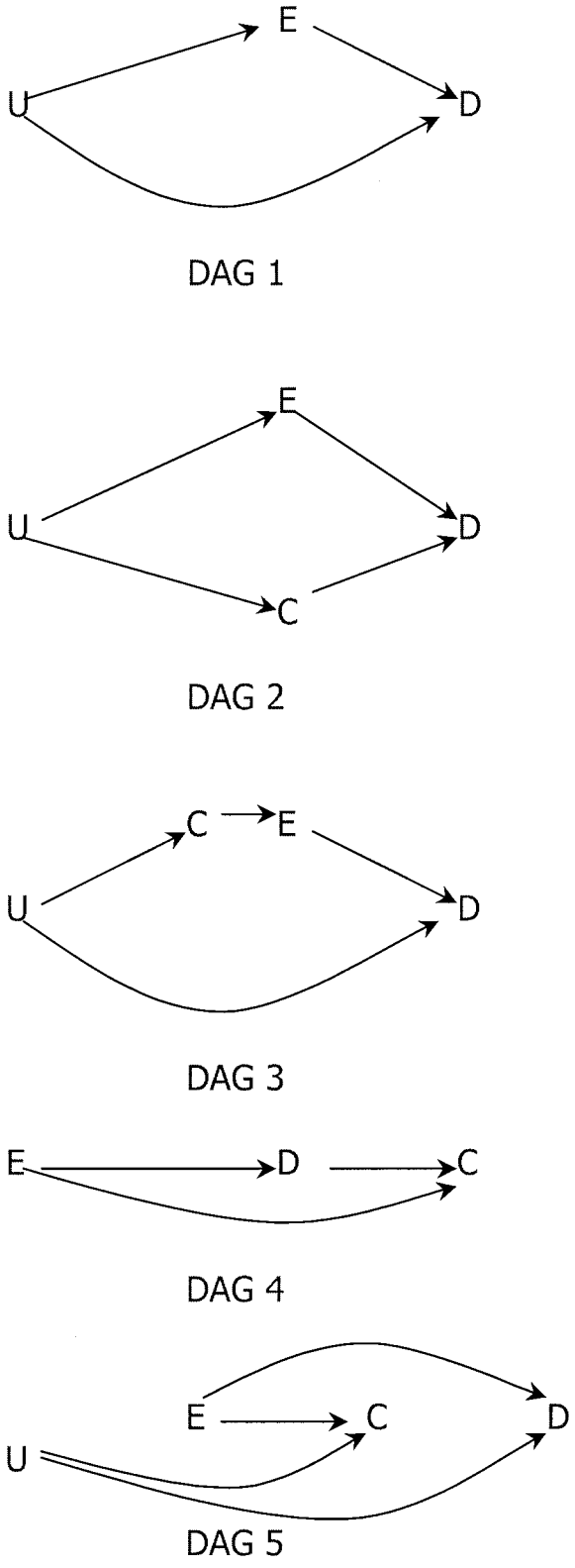


FIGURE 1. Directed acyclic graphs for the sections Causal Graphs and Using Causal Graphs to Check for Confounding, in Thought Experiment 1. D = disease; E = exposure; U = an unmeasured potential confounder; C = a measured potential confounder.

A direct cause of a variable V on the graph is called a parent of V, and V is called the parent's child. The variables that can be reached starting from V by following a sequence of directed arrows pointing away from V are the descendants of V. The ancestors of V are those variables with V as a descendant. We will assume V is a cause of each of its descendants but a direct cause only of its children (where direct is always relative to the other variables on the DAG). Thus, V is caused by all its ancestors, but only its parents are direct causes.

Consider DAG 3. C is a cause of D through the pathway $C \rightarrow E \rightarrow D$ but is not a direct cause. The intuition is that intervening and manipulating C will affect E, and the change in E will in turn affect D. If we intervene and set each subject's value of E to the same level (say, exposed), however, then additionally manipulating C will no longer affect the distribution of E and thus that of D. Hence, we say that C has no direct effect on D when controlling for (in the sense of intervening and physically controlling or setting) the variable E. Note, however, that U is both a direct cause of D and an indirect cause through the causal pathway $U \rightarrow C \rightarrow E \rightarrow D$.

Our causal DAGs are of no use unless we make some assumption linking the causal structure represented by the DAG to the statistical data obtained in an epidemiologic study. Recall that if a set of variables X is statistically independent of (that is, unassociated with) another set of variables Y conditional on a third set of variables Z, then, within joint strata defined by the variables in Z, any variable in X is unassociated with any variable in Y. For example, suppose all variables are dichotomous and the set Z consists of the two variables Z_1 and Z_2. Then conditional independence implies that the odds ratio between any variable in X and any variable in Y is 1 within each of the $4 = 2^2$ strata of Z: $(Z_1,Z_2) = (0.0)$, $(Z_1,Z_2) = (0,1)$, $(Z_1,Z_2) = (1,0)$, and $(Z_1,Z_2) = (1,1)$. The following so-called causal Markov assumption (CMA) links the causal structure of the DAG with various statistical independencies.

Causal Markov Assumption

On a causal graph, any variable that is not caused by a given variable V will be independent of V conditional on the direct causes of V.

Recall that the descendants of a variable V are those variables causally affected by V and that the parents of V are the variables that directly cause V. It follows that the CMA is the assumption that V is independent of its nondescendants conditional on its parents.

Example 1

On DAG 1, suppose that the arrow from E to D were absent so that neither E nor D causes the other. U represents all unmeasured common causes of E and D. Because U is the only parent of D, and E is not a descendant of D, the CMA implies that E and D are unassociated (that is, independent) given (that is, within strata of) U. That is, two variables that are not

causally related are independent conditional on their common causes.

It turns out that the CMA logically implies additional statistical independencies. Specifically, CMA implies that a set of variables X is conditionally independent of another set of variables Y given a third set of variables Z, if X is "d-separated" from Y given Z on the graph, where "d-separation,"[10,11] described below, is a statement about the topology of the graph.

To describe d-separation, we first need to define the "moralized ancestral" graph generated by the variables in X, Y, and Z. In the following, a path between 2 variables is any unbroken sequence of edges (regardless of the directions of any arrows) connecting the two nodes.

The moralized ancestral graph generated by the variables in X, Y, and Z is formed as follows:[11]

1. First, remove from the DAG all nodes (and corresponding edges) except those contained in the sets X, Y, and Z and their ancestors.
2. Next, connect by an undirected edge every pair of nodes that both (a) share a common child and (b) are not already connected by a directed edge.

The graph is referred to as "moralized," because, in step 2, we marry (connect) all unmarried (unconnected) parents of a common child.

X is d-separated from Y given Z if and only if on the moralized ancestral graph generated by X, Y, and Z, any path from a variable in X to a variable in Y intercepts (that is, goes through) some node in Z.

If X and Y are not d-separated given Z, we say they are d-connected given Z. Note that if there are no paths connecting variables in X to variables in Y on the moralized ancestral graph, then X and Y are d-separated.

To check for a crude (that is, unconditional or marginal) association, we make Z the empty set. It is crucial that one perform step 1 before step 2 when forming the moralized ancestral graph.

Example 2

Consider causal graph DAG 4. Note that *E* and C can have no common cause, because, if they did, that common cause would have to be represented on the graph. Now, assume there is no arrow from *E* to *D* so *E* does not cause *D*. Then, *E* and *D* are marginally independent (that is, have a crude odds ratio of 1). This statement follows either from the CMA or from the fact that *E* is d-separated from *D* given Z equal to the empty set. Specifically, in step 1 of the moralized graph algorithm, C and the arrows pointing into it are removed from the graph so that in step 2, *E* and *D* have no children. Thus, there is no path linking *E* and *D* on the moralized *ancestral* graph, so they are d-separated. This example tells us that two causally unrelated variables without a common cause are marginally unassociated (that is, independent).

In contrast, *E* and *D* are not d-separated given C. To see this, note that upon identifying Z as C, C is no longer removed in step 1 of the algorithm. Hence, in step 2, *E* and *D* have to be connected by an edge because they have a common child C. Hence, *E* and *D* are d-connected given C, because there is a direct edge between them in the moralized *ancestral* graph that does not intercept C. This example tells us that if we condition on a common effect C of two independent causes *E* and *D*, we "usually" render those causes conditionally dependent. For instance, if we know a subject has the outcome C (that is, we condition on that fact) but does not have the disease *D*, then it usually becomes more likely that the subject has the exposure *E* (because we require some explanation for his or her having C). That is, among subjects with the outcome C, *E* and *D* are "usually" negatively associated (have an odds ratio less than 1).

The reason we included the word "usually" in the above is that although CMA allows one to deduce that d-separation implies statistical independence, it does not allow one to deduce that d-connection implies statistical dependence. However, d-connected variables will generally be independent only if there is an exact balancing of positive and negative causal effects. For example, in DAG 3, *U* is a parent of and thus not d-separated from *D*. Yet if the direct effect of *U* on *D* is equal in magnitude but opposite in direction to the effect of *U* on *D* mediated through the variables C and *E*, then *U* and *D* would be independent, even though they are d-connected. Because such precise fortuitous balancing of effects is highly unlikely to occur, we shall henceforth assume that d-connected variables are associated.[6,7]

Using Causal Graphs to Check for Confounding

We can use causal graphs and d-connection to check for confounding as follows. First, suppose, as on DAGs 1-5, *E* is not an indirect cause of *D*. We begin by pretending that we know that exposure has no causal effect on the outcome *D* by removing just those arrows pointing out of exposure necessary to make *D* a nondescendant of *E*. If, under this causal null hypothesis, (1) *E* and *D* are still associated (that is, d-connected), then obviously the association does not reflect causation, and we say that the *E-D* association is confounded, and (2) if *E* and *D* are associated (d-connected) conditional on (that is, within levels) of Z, we say there is confounding for the *E-D* association within levels (strata) of Z. For example, the existence of an unmeasured common cause *U* of *E* and *D* as in DAG 1 will make *E* and *D* associated under the causal null (because *E* and *D* will be d-connected). If data on *U* have not been recorded for data analysis, confounding is intractable and we cannot identify the causal effect of *E* on *D*. If data on *U* are available, however, the conditional associations $OR_{ED|U}$ are unconfounded and will represent the causal effect of *E* on *D* within strata of *U*, that is, $OR_{ED|U} = OR_{causal,ED|U}$ at each level of *U*. This relation reflects the fact that under the causal null hypothesis of no arrow from *E* to *D*, I showed in Example 1 that *E* and *D* are independent (d-separated) given *U*. Furthermore, suppose, as has been assumed, that we have not conditioned on a variable lying on a casual pathway from *E* to *D*; then it is a general result that if *E* is a time-independent exposure

and *E* and *D* are (conditionally) independent under the causal null, then, under the causal alternative, the (conditional) association between *E* and *D* will reflect the (conditional) causal effect of *E* on *D*.[1]

Next we consider graphs 2 and 3, in which the variable C has been measured. Thus, in DAG 3, *U* remains an unmeasured common cause of *E* and *D*, although it is not a direct cause of *E*. It follows that, in both DAGs 2 and 3, the marginal association OR_{ED} is confounded, because *E* and *D* will be marginally associated (that is, d-connected) even under the causal null. However, the unmeasured variable *U* will not function as a common cause of *E* and *D* within strata of C because under the causal null *E* and *D* are d-separated given C. Thus, stratifying on C in the analysis will control confounding and $OR_{ED|C} = 1$ and $OR_{ED|C} = 0$ will represent the causal effect of *E* on *D* within strata of C. The variable *U* in DAGs 2 and 3 is referred to as a causal confounder, because it is a common cause of *E* and *D*. DAG 3 shows that we can control confounding due to a causal confounder *U* by stratifying on a variable C that itself is not a cause of *D*. Note, however, that C is an independent (but noncausal) risk factor for *D* in the sense that C and *D* are associated (d-connected) within strata of *E*.

Consider next DAG 4. There are no unmeasured common causes of *E* and *D*. As discussed in Example 2 above, under the causal null hypothesis of no arrow from *E* to *D*, *E* and *D* will be independent. It follows that the marginal association OR_{ED} is unconfounded and represents the causal effect of *E* on *D*. In contrast, the conditional association $OR_{ED|C}$ will be confounded and thus will not be equal to the causal effect of *E* on *D* within strata of C, because we showed in Example 2 that, under the causal null of no arrow from *E* to *D*, *E* and *D* will be conditionally associated within strata of C. This example shows that conditioning on a common effect C of *E* and *D* introduces confounding within levels of C. This example also shows why, to check for confounding, we remove from the graph just those arrows necessary for the outcome *D* to be a nondescendant of *E*; had we removed all arrows pointing out of *E* (including that into C) we would not have recognized that conditioning on C would cause confounding within levels of C.

An extension of this last example provides an explanation of the well-known adage that one must not adjust for variables affected by treatment. To see why, consider DAG 5, in which the exposure *E* has a direct causal effect on C, and C and *D* have an unmeasured common cause *U*. Under the causal null with the arrow from *E* to *D* removed, *E* and *D* will be d-separated and thus unassociated. Thus, the marginal association OR_{ED} will be unconfounded and represent causation. Nevertheless, the conditional associations $OR_{ED|C=1}$ and $OR_{ED|C=0}$ will be confounded and thus biased for the conditional causal effect within levels of C. This situation reflects the fact that, under the causal null, *E* and *U* will be associated once we condition on their common effect C. Thus, because *U* itself is correlated with *D*, *E* and *D* will be conditionally associated (that is, d-connected) within levels of C. Note the fact that the analysis stratifying on C was confounded even under the causal null proves that adjusting for a variable C affected by treatment can lead to confounding and bias even when C is not an intermediate variable on any causal pathway from exposure to disease.

Finally, suppose, on a causal graph, *E* is an indirect cause of *D* through a directed path $E \rightarrow C \rightarrow D$ so that, among those with $C=c$, the net (overall) effect $OR_{causal,ED|C=c}$ differs from the direct effect of *E* on *D*. We can still graphically test for confounding as described above, except that, now, regardless of our test results, we must never conclude that $OR_{ED|C=c}$ equals $OR_{causal,ED|C=c}$ for any variable C on a causal pathway from *E* to *D*.

With this background, we are ready to justify the answers given above.

Justifications of Answers

Case-Control Study

We first argue that the causal graph representing our case-control study is DAG 6 (Figure 2). By assumption, we need not worry about unmeasured preconception confounders. Furthermore, we know that if there is an arrow between *E* and *D*, it must go from *E* to *D* because the medical records were created in the first trimester, before the development of the second trimester congenital defect. Also, actually taking a medicine will be a cause of a woman reporting that she took a medicine; hence, the arrow from *E* to *E**. Finally, because a woman's self-report, *E**, is obtained after her child's birth, the defect *D* will be a cause of *E**, if, as is likely, mothers whose children have a congenital defect are more prone to recall their medications than are other mothers. We can use the data to confirm the existence of an arrow from *D* to *E**, because otherwise *E** and *D* would be independent (d-separated) within levels of *E*. But one can check from Table 1 that among subjects with $E = 1$, *D* and *E** are associated ($OR_{DE^*|E=1} = 0.3$), so misclassification is differential. DAG 6 is isomorphic to DAG 4 with *E** playing the role of C. Thus, as in DAG 4, we conclude that the marginal association $OR_{ED} = 1.7$ is causal but the conditional association $OR_{ED|E^*} = 3$ will differ from the conditional causal effect $OR_{causal,ED|E^*}$. Mistakenly interpreting $OR_{ED|E^*} = 3$ as causal could in principle lead to poor public health decisions, as would occur if a cost-benefit analysis determines that a conditional causal odds ratio of 2.9 is the cutoff point above which the risks of congenital malformation outweigh the benefits to the mother of treatment with *E*.

Finally, a possibility that we have not considered is that those mothers who develop, say, a subclinical infection in the first trimester are at increased risk both of a second trimester congenital malformation and of worsening arthritis, which they may then treat with the drug *E*. In that case, we would need to add to our causal graph an unmeasured common cause *U* (subclinical infection) of both *E* and *D* that represents subclinical first trimester infection, in which case even OR_{ED} would be confounded.

Prospective Cohort Study

In the prospective cohort study, sufficient information is given so that we know there is no confounding by unmeasured pre-employment factors. Yet, as noted above, E^* is associated with D given E. Now clearly E^*, which is a measure of the air-level of radon in mines, cannot itself directly cause death other than through its effect on a subject's actual pulmonary radon exposure E, so that there cannot be a direct arrow from E^* to D. Nevertheless, because E^* was measured before death, D cannot be a cause of E^* either. Furthermore, we are given that there is no arrow from any unmeasured confounder into E, because, although physical exertion is a cause of the pulmonary dose E, it is not a cause of D. The most reasonable explanation for these facts is that E^* is a surrogate for some other unmeasured adverse causal exposure in the mine (say silica). Thus, we might consider the causal graph shown in DAG 7. In this figure, *MINE* represents the particular mine in which the subject works. It is plausible that mines with high levels of radon may have low levels of silica-bearing rock (because silica-bearing rock is not radioactive). Therefore, E^* and *SILICA* will be negatively correlated. If DAG 7 is the true causal graph (with *MINE* and *SILICA* being unmeasured variables), then under the causal null hypothesis in which the arrow from E to D is removed, E and D will still remain correlated because *MINE* is an unmeasured common cause of E and D but, by d-separation, E and D will be independent conditional on E^*. Thus, OR_{ED} is confounded; however, $OR_{DE|E^*} = 3$ equals the causal effect $OR_{causal,DE|E^*}$ of exposure on disease within strata of E^*0.3. In contrast, the conditional association $OR_{E^*D|E} = 0.3$ represents not a protective effect of E^* on D but rather the negative correlation between E^* and *SILICA* conjoined with the adverse causal effect of *SILICA* on D. DAG 7, however, probably does not tell the whole story. One would expect that physical exertion is a direct cause of a worker's actual (unrecorded) silica dose. Thus, physical exertion is an unmeasured common cause of E and D, even when we condition on E^*, precluding unbiased estimation of the causal effect of E on D.

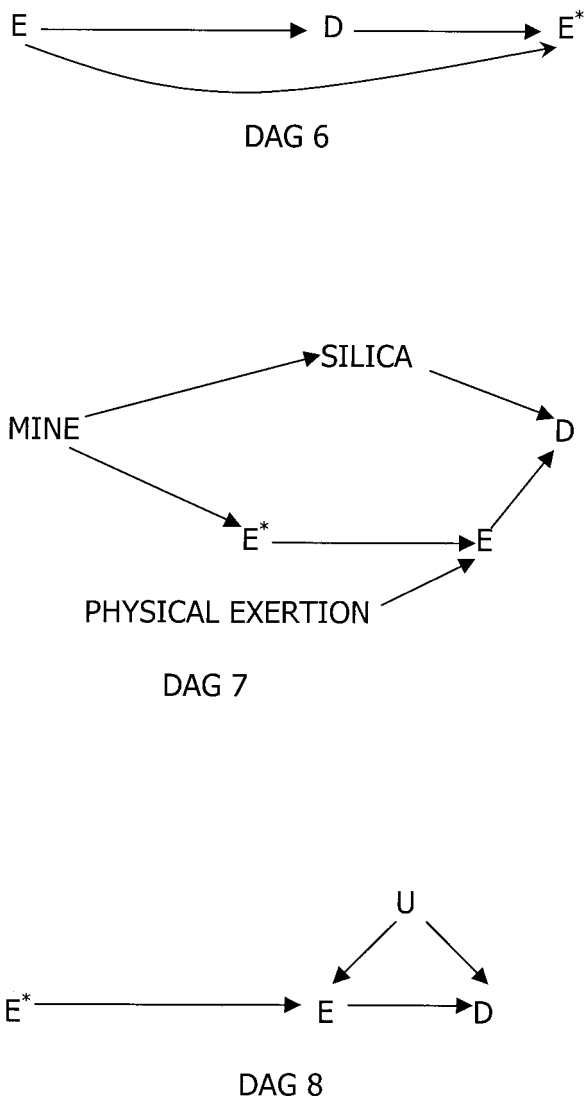


FIGURE 2. Directed acyclic graphs (DAG) for the Justifications of Answers section in Thought Experiment 1. D = disease (in DAG 6, D = congenital defect in offspring); E = exposure; E* = a misclassified version of E; U = an unmeasured common cause of E and D, such as underlying health status.

Randomized Clinical Trial

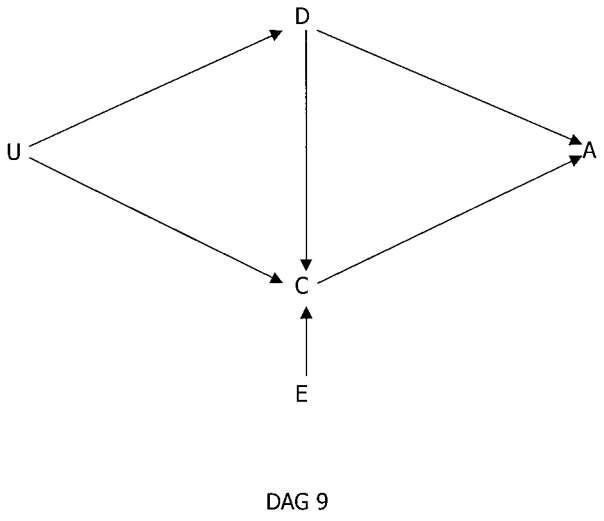
The study is a typical randomized trial with noncompliance and is represented by the causal graph in DAG 8.[12] Because E^* was randomly assigned, it has no arrows into it. Given assignment, however, both the decision to comply and the outcome D may well depend on underlying health status U. E^* has no direct arrow to D, because, by assumption, E^* causally influences D only through its effect on E. We observe that under the causal null in which the arrow from E to D is removed, E and D will be associated (d-connected) owing to their common cause U both marginally and within levels of E^*. Hence, both OR_{ED} and $OR_{ED|E^*}$ are confounded and have no causal interpretation. Under this causal null, however, E^* and D will be independent, because they have no unmeasured common cause. Hence, we can test for the absence of an arrow between E and D (that is, lack of causality) by testing whether E^* and D are independent. This test amounts to the standard intent-to-treat analysis of a randomized trial. Thus, even in the presence of nonrandom noncompliance as a result of U, an intent-to-treat analysis provides for a valid test of the causal null hypothesis that E does not cause D. Because $OR_{E^*D} = 0.5$ in our data, we conclude that we can reject the causal null and that E protects against D in at least some patients. Now, OR_{E^*D} represents the effect of assignment to a low-fat diet on the outcome. Owing to noncompliance, this measure in general will differ from the causal effect $OR_{causal,ED}$ of actually following a low fat diet. Indeed, the magnitude $OR_{causal,ED}$ of the causal

effect of *E* in the study population is not identified (that is, estimable), and one can only compute the bounds for it. Finally, note that the conditional association $OR_{E^*D|E} = 0.3$ also fails to have a causal interpretation. This conclusion reflects the fact that under the causal null of no arrow from *E* to *D*, *E** and *D* will be conditionally associated within levels of *E*, because *E* is a common effect of both *E** and *U*, and *U* is a cause of *D*.

Thought Experiment 2: Postmenopausal Estrogens and Endometrial Cancer

Consider causal DAG 9 with *D* being endometrial cancer, C being vaginal bleeding, *A* being ascertained (that is, diagnosed) endometrial cancer, *E* being postmenopausal estrogens, and *U* being an unmeasured common cause of endometrial cancer and vaginal bleeding (Figure 3). For simplicity, we assume that our diagnostic procedures have 100% sensitivity and specificity. So, every woman with *D* who receives a diagnostic test will be successfully ascertained, as is represented by the arrow from *D* to *A*. There may, however, be many women with endometrial cancer who have not had a diagnostic procedure and thus remain undiagnosed.

The absence of an arrow from *E* to *D* represents the Horowitz and Feinstein[2] null hypothesis that estrogens do not cause cancer. The arrow from *E* to C indicates that estrogens cause vaginal bleeding. The arrows from C to *A* indicate that vaginal bleeding leads to endometrial cancer being clinically diagnosed. The arrow from *D* to C indicates that endometrial cancer can cause vaginal bleeding. The arrows from *U* to *D* and C indicate that some unknown underlying uterine abnormality *U* independently leads to both uterine bleeding and cancer. We will also consider subgraphs of DAG 9 with various arrows removed.



DAG 9

FIGURE 3. DAG for Thought Experiment 2. D = endometrial cancer; A = ascertained endometrial cancer; C = vaginal bleeding; E = exogenous estrogens; U = an unmeasured common cause of D and C.

There will be ascertainment bias whenever the arrow from C to *A* is present, because then, among women with endometrial cancer, those who also have vaginal bleeding are more likely to have their cancer diagnosed.

Furthermore, *D* and C will be associated (d-connected) in the source population whenever either (1) endometrial cancer causes vaginal bleeding so that the arrow from *D* to C is present or (2) *U* is a common cause of cancer and bleeding so that the arrows from *U* to *D* and from *U* to C are present.

Now consider a case-control study in which we find each clinically diagnosed case of endometrial cancer *D* in a particular locale and select as a control a random age-matched woman yet to be diagnosed with endometrial cancer. If we let *b* be the number of discordant pairs with the case exposed and *c* be the number of discordant pairs with the control exposed, *b*/*c* is the Mantel-Haenszel odds ratio (MH OR).

The MH OR is biased (that is, converges to a value other than 1) under the null hypothesis of no estrogen effect on endometrial cancer if and only if there is ascertainment bias. To see this, note that, under this design, the MH OR will converge to 1 (that is, be unconfounded) if and only if *A* (diagnosed cancer) is unassociated with the exposure *E*. But *E* and *A* are associated (d-connected) if and only if there is an arrow from C to *A*.

To adjust for vaginal bleeding, we might consider a second bleeding-matched design in which we additionally match controls to cases on the presence of vaginal bleeding in the month before the cases' diagnosis. Under this design, whether or not ascertainment bias is present, the bleeding-matched MH OR is biased away from 1 if and only if endometrial cancer *D* and vaginal bleeding C are associated (d-connected), owing to an unmeasured common cause *U* or to *D* causing C or to both. This result follows by noting that the bleeding-matched MH OR is 1 if and only if *A* is independent of (d-separated from) *E* conditional on C. But, *A* is d-separated from *E* given C if and only if *D* and C are unassociated. It follows that we have given a graphical proof of the well-known result that one cannot control for ascertainment bias by stratification on determinants of diagnosis if these determinants are themselves associated with disease.[5]

Combining the results, we can conclude that in the presence of both a vaginal bleeding-endometrial cancer association and ascertainment bias, the MH OR and the bleeding-matched MH OR are both biased.

It is now clear why there was a controversy: On biological and clinical grounds, it was believed that endometrial cancer caused vaginal bleeding and that vaginal bleeding led to the ascertainment of undiagnosed cancer. Thus, one could not validly test the Horowitz and Feinstein[2] null hypothesis whether or not one controlled for the determinant of ascertainment bias (that is, vaginal bleeding) in the analysis. We note that Greenland and Neutra,[5] Hutchison and Rothman,[3] and Jick *et al*[4] reach conclusions identical to ours. Our contribution is to demonstrate how quickly and essentially

automatically one can reach these conclusions by using causal graphs.

Discussion

If every pair of variables had one or more unmeasured common causes, then all exposure-disease associations would be confounded. I believe that, in an observational study, every two variables have an unmeasured common cause, and thus there is always some uncontrolled confounding. Thus, when, as in our examples, one considers causal graphs in which certain pairs of variables have no unmeasured common causes, this situation should be understood as an approximation. Of course, in an observational study, we can never empirically rule out that such approximations are poor, as there may always be a strong unmeasured common cause of which we were unaware. For example, in the case-control study of our first thought experiment, those without sufficient subject matter expertise would not have had the background needed to recognize the possibility that a subclinical first trimester infection might be a common cause of exposure and the outcome. As epidemiologists, we should always seek highly skeptical subject-matter experts to elaborate the alternative causal theories needed to help keep us from being fooled by noncausal associations.

References

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.
2. Horowitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med 1978;299:1089–1094.
3. Hutchison GB, Rothman KJ. Correcting a bias? N Engl J Med 1978;299:1129–1130.
4. Jick H, Watkins RN, Hunter JR, Dinan BJ, Madsen S, Rothman KJ, Walker AM. Replacement estrogens and endometrial cancer. N Engl J Med 1979;300:218–222.
5. Greenland S, Neutra R. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. J Chron Dis 1981;34:433–438.
6. Pearl J, Verma T. A theory of inferred causation. In: Allen JA, Fikes R, Sandewall E, eds. Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference. San Mateo, CA: Morgan Kaufmann, 1991:441–452.
7. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. New York: Springer Verlag; 1993.
8. Pearl J. Causal diagrams for empirical research. Biometrika 1995;82:669–688.
9. Pearl J, Robins JM. Probabilistic evaluation of sequential plans from causal models with hidden variables. In: Uncertainty in Artificial Intelligence: Proceedings of the 11th Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1995;444–453.
10. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann, 1988.
11. Lauritzen SL, Dawid AP, Larsen BN, Leimar HG. Independence properties of directed Markov fields. Networks 1990;20:491–505.
12. Balke A, Pearl J. Bounds on treatment from studies with imperfect compliance. J Am Stat Assoc 1997;92:1171–1176.