

We consider an epidemiologic thought experiment to make the point that the choice of an appropriate etiologic (causal) analysis depends as much on the design of the study and background subject-matter knowledge as on the data.

1 Thought Experiment

Consider the data given in Tables 1 and 2 .

E is a correctly classified exposure whose causal effect on D I would like to ascertain.

E^* is a misclassified version of E .

Data on E , E^* , and D are available on all study subjects (unusual).

Sampling variability can be ignored.

I will report all associations on an odds ratio scale. This choice is dictated by the fact that in study (a) below, the only estimable population association measures are odds ratios.

Recall for any variable Z , the exposure-disease causal odds ratio among the subset of subjects with Z being z is

$$\begin{aligned} & OR_{ED|Z=z} \\ = & \frac{pr [D = 1|Z = z, E = 1] / pr [D = 0|Z = z, E = 1]}{pr [D = 1|Z = z, E = 0] / pr [D = 0|Z = z, E = 0]} \\ = & \frac{pr [D = 1|Z = z, E = 1] pr [D = 0|Z = z, E = 0]}{pr [D = 1|Z = z, E = 0] pr [D = 0|Z = z, E = 1]} \end{aligned}$$

Table 1a

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

$$OR_{E^*D|E=1} = OR_{E^*D|E=0} = 9$$

Table 1b

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

$$OR_{E^*D|E=1} = OR_{E^*D|E=0} = .3$$

Table 1c

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

$$OR_{ED|E^*=1} = OR_{ED|E^*=0} = 3$$

Table 2a		
	$E = 1$	$E = 0$
$D = 1$	380	220
$D = 0$	800	800
$OR_{ED} = 1.73$		

Table 2b		
	$E^* = 1$	$E^* = 0$
$D = 1$	200	400
$D = 0$	800	800
$OR_{ED} = .5$		

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:

Questions: Can you 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?

(a) **Case Control Study**: Suppose the data arose from a case-control study of the effect of a particular non-steroidal 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 one month post-partum by maternal self-report.

The data E were obtained from comprehensive accurate medical records of first trimester medications.

All relevant pre-conception confounders and other drug exposures were controlled by stratification.

The data in Table 1 are taken from a particular stratum.

Note misclassification is differential since $OR_{E^*D|E=1} = OR_{E^*D|E=0} = .3 \neq 1$.

(b) 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 (below) threshold exposure to radon as measured by lung dosimetry.

Each miner was also assigned an estimated radon exposure E^* based on 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 six months of physical exertion at age 25 has no independent effect on later mortality.

(c) **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 .

1.1 Causal Contrasts

To determine which association measure is most likely causal, we need a formal definition of causal effects.

Causal effects are best expressed in terms of counterfactual variables.

Let the variable $D_{e=1}$ denote a subject's outcome if exposed and $D_{e=0}$ denote a subject's outcome if unexposed.

For a given subject, the causal effect of treatment, measured on a difference scale, is $D_{e=1} - D_{e=0}$.

If a subject is exposed ($E = 1$), the subject's observed outcome D equals $D_{e=1}$, and $D_{e=0}$ is unobserved.

If $E = 0$, D equals $D_{e=0}$, and $D_{e=1}$ is unobserved.

Let $pr(D_{e=1} = 1)$ and $pr(D_{e=0} = 1)$ respectively be the probability that $D_{e=1}$ is equal to 1 and $D_{e=0}$ is equal to 1 where probabilities refer to proportions in a large, possibly hypothetical, source population.

Then, the exposure-disease causal odds ratios is

$$\begin{aligned}
 & OR_{causal,ED} \\
 = & \frac{pr(D_{e=1} = 1) / pr(D_{e=1} = 0)}{pr(D_{e=0} = 1) / pr(D_{e=0} = 0)} \\
 = & \frac{pr(D_{e=1} = 1) pr(D_{e=0} = 0)}{pr(D_{e=1} = 0) pr(D_{e=0} = 1)}
 \end{aligned}$$

. For any variable Z , the exposure-disease causal odds ratio among the subset of subjects with Z being z is

$$\begin{aligned}
 & OR_{causal,ED|Z=z} \\
 = & \frac{pr[D_{e=1} = 1|Z = z] pr[D_{e=0} = 0|Z = z]}{pr[D_{e=1} = 0|Z = z] pr[D_{e=0} = 1|Z = z]}
 \end{aligned}$$

1.2 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 (a), 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 $OR_{ED|E^*} = 3$ is biased in the sense that it fails to equal the causal effect $OR_{causal,ED|E^*}$ of exposure on disease among subjects within a particular stratum of E^* .

In the prospective cohort study (*b*), exposure is likely beneficial and the best parameter choice is the conditional odds ratio $OR_{DE|E^*} = 3$.

In the randomized trial (*c*), exposure is likely beneficial and the best parameter choice may be the crude E^*D association $OR_{E^*D} = .5$, although it is likely that this underestimates the true benefit of exposure.

1.3 Justifications of Answers

(a). We argue that the causal graph representing our case-control study is DAG 6.

By assumption, we need not worry about unmeasured preconception confounders.

Further, we know that if there is an arrow between E and D , it must go from E to D since 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, since 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 $OR_{DE^*|E=1} = .3$, 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 cut-off 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:

Mothers who develop a sub-clinical infection in the first trimester are both at increased risk 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 sub-clinical first trimester infection, in which case even OR_{ED} would be confounded.

(b). 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.

Further, 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 (since 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^* .

In contrast, the conditional association $OR_{E^*D|E} = .6$ 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 .*

(c). The study is a typical randomized trial with non-compliance. Under the exclusion restriction of no direct effect of assigned treatment on D , it is represented by the causal graph in DAG 8.

Since 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 , since, 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, since they have no unmeasured common cause. Hence we can test for the absence of an arrow between E and D (i.e., 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 non-random non-compliance due to U , an intent-to-treat analysis provides for a valid test of the causal null hypothesis that E does not cause D . Since $OR_{E^*D} = .5$ in our data, we conclude we can reject the causal null and conclude that E protects against D in at least some patients.

Nonetheless, OR_{E^*D} represents the effect of assignment to a low fat diet on the outcome.

Owing to non-compliance, this measure in general will differ from the population 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 (i.e., estimable), and one can only compute the bounds for it.

Finally, note that the conditional association $OR_{E^*D|E} = .3$ 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 , since E is a common effect of both E^* and U , and U is a cause of D .

