On effect modification and its applications

In a deterministic universe component causes join hands to form a sufficient cause of an outcome. For example, the mutated gene for phenylalanine hydroxylase awaits the arrival of phenylalanine in the diet to complete a sufficient cause of mental retardation. When both are present in the baby's body, the devastating outcome is inevitable—or so you may think.

That viewpoint takes us to the deterministic idea of interaction: cause A (a mutated gene) interacts with cause B (amino acid intake) to bring about the effect C (mental retardation). Paradoxically perhaps, the word interaction also found home in frequentist statistics (interaction models), even though frequentist statistics is founded on chance-type probabilities, which resonate with indeterminism.

Believers in determinism could have ended the story here, but a related term—effect modification—disturbed their flawless image of interacting causes. Instead of assuming that a mutated gene (cause A) interacts with phenylalanine intake (cause B) to form a sufficient cause of mental retardation, we may assume that the risk of mental retardation, given genotype (variable A), varies according to phenylalanine intake (variable B)—and vice versa. For instance, the intake of phenylalanine may have null or negligible effects on mental retardation when the gene is normal, but may have strong effects otherwise. A new statistical model is not needed, either. The so-called interaction model may also be described as a model of effect modification.

Faithful to interacting causes, deterministic writers struggle to inject some meaning into the term "effect modification". Some of them downgraded it to "effect measure modification"; others confused it with "association modification"; and others tried to subsume it under the "consistency thingamajig" (commentary here). I have no idea why they bother. Subscribers to determinism should simply declare the term superfluous, because in a deterministic universe there are no modifiers, only "interactors"—the components of a sufficient cause.

Just in case determinism is false and sufficient causes do not exist, let's review some aspects of effect modification and explore its helpful consequences.

First, the magnitude of effect modification (zero is magnitude, too) depends on the scale on which effects are measured, and the question of which scale is preferred under indeterminism, if any, awaits a solid analysis. The available literature is small and not particularly illuminating. It is possible that mathematical tools cannot fully reveal some aspects of causal reality, perhaps because the axioms of math are not sufficient for the task.

Second, in methodology we often resort to a dichotomy between a non-null effect and a null effect, or between the presence and absence of effect modification. In deep science, however, the interesting questions are quantitative: What is the effect size? What is the magnitude of effect modification? Neither effects nor effect modification should be reduced to all-or-none phenomena. Again, methodological pieces (like this one) are exempt.

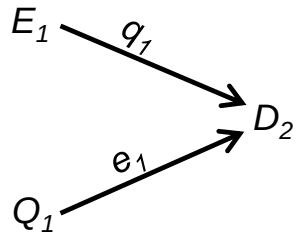
Third and most important, I used to think that effect modification only requires two causal variables and some shared outcome.[1] I have now accepted the premise that *only coinciding variables may be effect modifiers*. Blood pressure now (Q_1) may modify the effect of blood glucose now (E_1) on stroke tomorrow (D_2), but blood pressure an hour ago (Q_0) does not modify any effect of blood glucose now (E_1). Just as an unmodified effect operates between two time points ($E_1 \rightarrow D_2$), so does a parallel modifier, Q_1, of that effect ($Q_1 \rightarrow D_2$). A modifier simply allows for two or more causal parameters within a single arrow. For example:

$$E_1 \rightarrow D_2 | Q_1=1 \quad \neq \quad E_1 \rightarrow D_2 | Q_1=0$$

Read: the effect of E_1 on D_2 when Q_1=1 is not identical to the effect of E_1 on D_2 when Q_1=0 (given $Q_1 \rightarrow D_2$)

Figure 1 shows intuitive notation for effect modification. The arrow is supplemented with a lower-case letter, indicating dependency of the causal parameter on the modifier's value at t=1. As shown in the figure, effect modification is a reciprocal phenomenon.[2]

Figure 1. Displaying effect modification



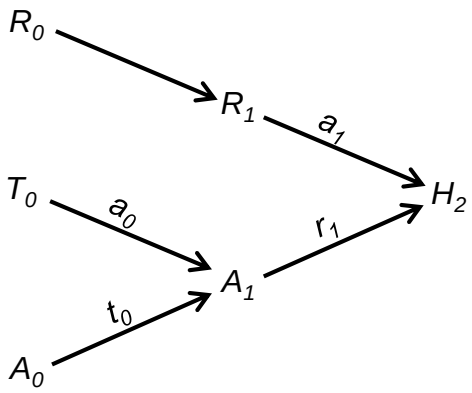
According to these premises, the literature on effect modification by an intermediary (e.g., $E_1 \rightarrow Q_2 \rightarrow D_3$, where Q_2 is called "a modifier") is a mathematical

exercise, rather than a description of causal reality—analogous to the math of "treatment regime", time-dependent confounders, and change variables.[3,4] Not every mathematical derivation that looks like science is science indeed.

Effect modification and treatment

Medical treatment often exploits the phenomenon of effect modification. Consider, for example, the symptom of heartburn (*H*), which may be caused by reflux (*R*) of acidic stomach fluid (*A*) into the esophagus, and may be treated by drugs (*T*) that raise the pH of stomach fluid. Figure 2 shows part of the causal structure for three time points. (The arrows $R_0 \rightarrow R_1$ and $A_0 \rightarrow A_1$ follow an axiom of causality.[3])

Figure 2. Causes and treatment of heartburn



R_1 and A_1 are effect modifiers: In particular, the higher the acidity of stomach fluid (lower pH)—the stronger the effect of reflux on heartburn. To alleviate heartburn, we may try to raise the pH of stomach fluid, and to that end we turn to the causes of A_1. One of these causes is treatment status (T_0).

But the treatment effect on stomach fluid acidity ($T_0 \rightarrow A_1$) is also grounded in effect modification. Coinciding with T_0 is the acidity level, A_0, which is another cause of A_1. As shown in Figure 2, T_0 and A_0 are effect modifiers. Specifically, when T_0=no treatment, the effect $A_0 \rightarrow A_1$ is strong: if A_0 takes the value "low pH", A_1 is also likely to take the value "low pH". In contrast, when T_0=proton pump inhibitor, the effect $A_0 \rightarrow A_1$ is weak, and the treatment, T_0, strongly affects A_1 in the desired direction: A_1 is likely to take the value "high pH" (even if A_0 took the value "low pH"). Recall that the lower the acidity of stomach fluid—the weaker the effect of reflux on heartburn.

To sum up, at least two modified effects underlie contemporary treatment of heartburn (and many other conditions).

Effect modification and causal paths

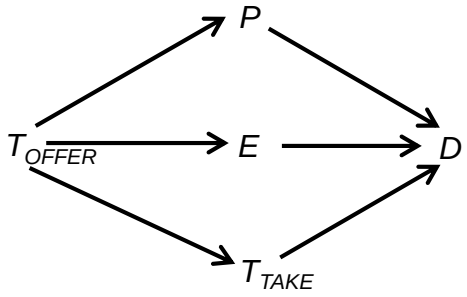
Quite often an arrow may be decomposed into several causal paths, some of which are specified through intermediary variables (eg, $E_0 \rightarrow X_1 \rightarrow D_2$) and others may be summarized by a single arrow ($E_0 \rightarrow D_2$), commonly miscalled "the direct effect".

I used to think that "partial effect"—the remainder effect after analytical exclusion of some causal path(s)—is a valid idea.[1] I have now accepted the premise that *partial effects do not exist*; their computation also crosses the boundary between causal reality and a mathematical exercise. To ask about the effect of weight at t=0 on stroke at t=2 —if blood glucose at t=1 would be fixed to 100mg/dL—is a human-made hypothetical, not causal reality. It has no place in indeterminism, and I am not sure that it fits well with determinism, either.

Although partial effects do not exist, we may still be able to nullify causal paths that are not of immediate interest—exploiting again, effect modification. Double-blinded trials provide an example.

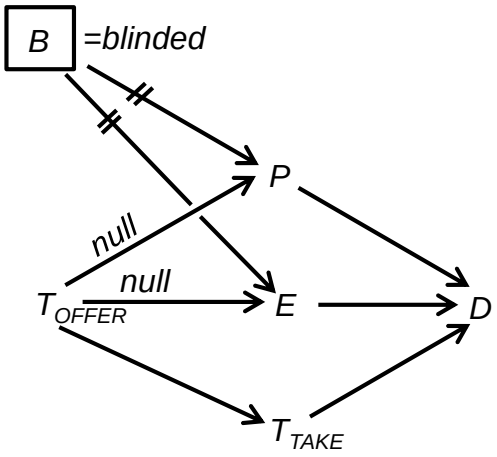
Figure 3 shows how the effect of the offered treatment (T_{OFFER}) on the outcome (*D*) may be decomposed into at least three paths: 1) through taking the treatment (T_{TAKE}); 2) through expectation of benefit (*E*); and 3) through subsequent treatment by the physician (*P*).

Figure 3. Causal paths for the offered treatment



Whenever a new treatment is tested, we are not interested in its effect through the patient's expectation or through subsequent treatment. We would like to estimate its effect when both of these paths are nullified, which requires modifiers of $T_{OFFER} \rightarrow E$ and $T_{OFFER} \rightarrow P$. And that is exactly what double blinding is supposed to achieve (Figure 4).

Figure 4. Nullified causal paths in a blinded trial

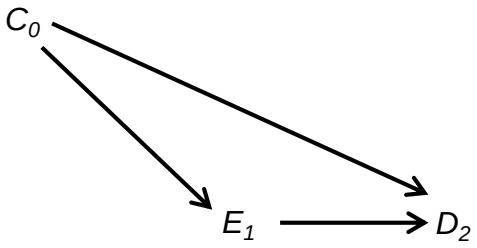


If neither the patient nor the physician knows what the patient is offered (*B=blinded*), the offered treatment should have a null effect on expectation of benefit (or harm) and on subsequent treatment. Therefore, the causal paths $T_{OFFER} \rightarrow E \rightarrow D$ and $T_{OFFER} \rightarrow P \rightarrow D$ would not contribute to the association between T_{OFFER} and D (Figure 4). Blinding is also expected to nullify some paths between the offered treatment and the analyzed outcome (not shown), and thereby reduce information bias. For instance, the offered treatment should not affect endpoint classification, if the classifier does not know which treatment was offered.

Effect modification and confounding bias

Confounding bias arises because the exposure (E_1) and the disease (D_2) share a cause (C_0)—a confounder (Figure 5). The association between E_1 and D_2 is accounted for not only by the causal path, $E_1 \rightarrow D_2$, but also by the confounding path, $E_1 \leftarrow C_0 \rightarrow D_2$.

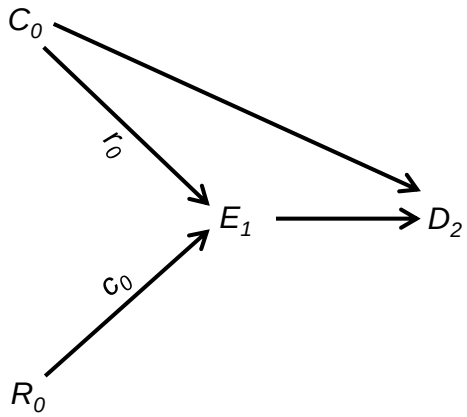
Figure 5. Confounding bias



A confounding path may be blocked by conditioning on the confounder, which dissociates the variable from both the exposure and the disease. But a confounding path may also be eliminated by finding a condition in which the confounder's effect on the exposure ($C_0 \rightarrow E_1$) is null. If the confounder C_0 does not affect the exposure, one segment of the confounding path is absent, and E_1 and D_2 are no longer associated through C_0.

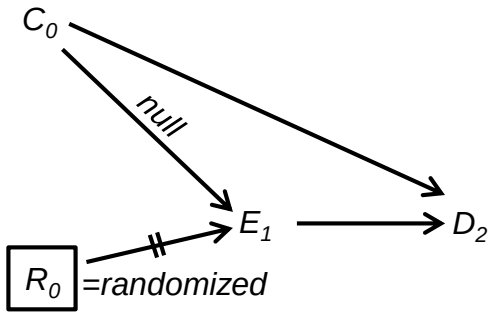
That $C_0 \rightarrow E_1$ can be null in some conditions, but not in others, implies effect modification by at least one other cause of E_1. A candidate modifier, R_0, may be the mechanism by which exposure status is offered (Figure 6).

Figure 6. Effect modification between a confounder and exposure assignment



Whenever exposure status is offered at will, we assume that C_0 has some effect on E_1, so confounding bias is present. If, however, exposure status were truly offered by a random process (R_0 = *randomized*) then C_0 has a null effect on E_1, by definition (Figure 7). In general, we assume that every confounder has a null effect on E_1, given randomization, which implies that all confounding paths have been eliminated. Modification of the confounders' effects on the exposure formally explains the well-known claim that confounding bias is absent from an intention-to-treat analysis of a randomized trial.

Figure 7. Deconfounding by randomization

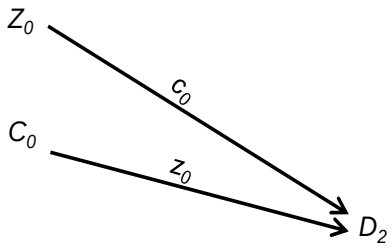


Reality, however, is more complex. The outcome of a random process, as displayed on a monitor or paper, affects the offered treatment, but is not synonymous with the actual offering of a treatment. Which makes room for confounding bias: the offered treatment

and the disease it may affect could still share a cause.[5] An intention-to-treat analysis of a randomized trial indeed reduces sources of confounding bias, but does not guarantee its absence.

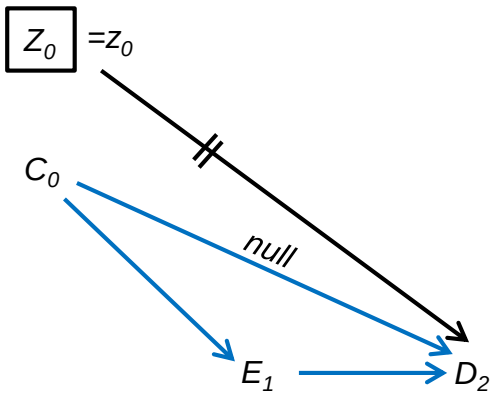
I wondered whether it is also possible to eliminate confounding by finding a parallel situation where the confounder's effect on the disease is modified by some cause of the disease (Figure 8).

Figure 8. Effect modification between the confounder and some cause of the disease



Although a parallel idea of randomized disease does not exist, the answer is, yes—if only theoretically. Some modifier (Z_0) might have a value (z_0) that nullifies the component of the causal association between C_0 and D_2 that complements the path $C_0 \rightarrow E_1 \rightarrow D_2$ (Figure 9). If we happen to restrict the modifier to that value—lucky conditioning—the confounding path no longer exists. Unfortunately, the associational condition is unpredictable, and luck is not counted among the methods of research.

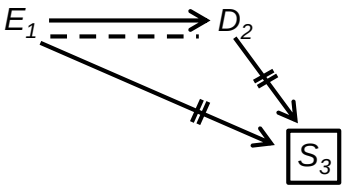
Figure 9. Theoretical deconfounding by conditioning



Effect modification and colliding bias

Unlike confounding bias, colliding bias can arise by several causal structures, the simplest of which requires a shared effect (S_3) of the exposure (E_1) and the disease (D_2). Following conditioning on S_3, a new component is often added to the association between E_1 and D_2 (Figure 10).

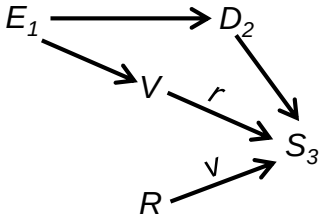
Figure 10. Colliding bias



In many examples S_3 is selection status, a binary variable that takes the value "selected for the study" or "not selected". Since only the selected people are eventually studied, conditioning on S_3 is inherent in research and would lead to colliding bias if coupled with the structure $E_1 \rightarrow S_3 \leftarrow D_2$. The colliding arrows themselves might arise through intermediaries that are beyond our control or through erroneous selection criteria.

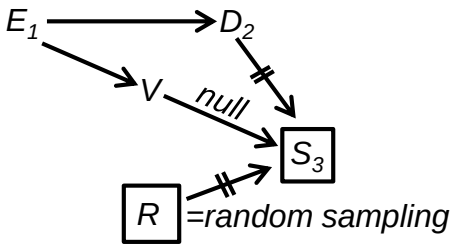
Although no study is free of selection criteria—a potential source of colliding bias—we may be able to prevent *some* colliding bias by finding a modifier of a colliding arrow into S_3 (Figure 11).

Figure 11. Effect modification between two causes of selection status, one of which is affected by the exposure



A candidate modifier is the sampling method (R). When sampling follows a random process, the path $E_1 \rightarrow V \rightarrow S_3$ may be nullified for *some* V (Figure 12), preventing colliding bias through that variable.

Figure 12. Preventing colliding bias by random sampling



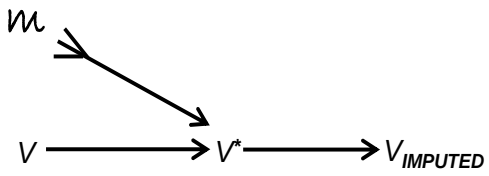
It is interesting to compare random sampling with randomization, two ideas that the novice tends to confuse. Randomization can remove confounding bias, whereas random sampling can prevent colliding bias. Randomization is expected to block all existing paths of confounding bias, whereas random sampling is expected to prevent some paths of colliding bias that would have otherwise arisen—say, by an erroneous selection rule. In both cases, however, the benefit is gained by conditioning on a modifier: the mechanism by which exposure status is offered (randomization), or the mechanism by which people are selected (random sampling).

Effect modification and information bias

Much of science is pursued in a parallel world of measured variables, or more precisely—"imputed variables": the values that your software analyzes when you click "run", "enter", or whatever. The causal paths that take us from the variable of interest (V) to its imputed version ($V_{IMPUTED}$) are long and complex,[6] and it is unclear which intermediary, if any, deserves the title "the measurement of V". We will ignore that complexity, however, and consider an arbitrary intermediary (V^*). In many cases V^* is a cause of the analyzed variable (Figure 13), although there are other methods to arrive at $V_{IMPUTED}$.[6]

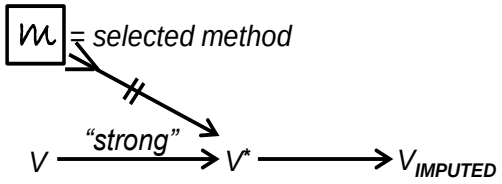
V^* has many causes, one of which is obviously V, the variable that is being measured ($V \rightarrow V^*$). Other causes are technical variables ($M1$, $M2$, $M3$….) that denote possible measurement methods. Figure 13 depicts them collectively as $\mathcal{M}$; their many effects on V^* are displayed by a single arrow with a three-pronged tail.

Figure 13. Causes of $V_{IMPUTED}$



To choose a method for measuring V^* is to condition on the variables that make up $\mathcal{M}$. But we do not condition arbitrarily. We prefer a method (equipment, protocol, training, and so on) for which the effect $V \rightarrow V^*$ is assumed to be strong rather than weak (Figure 14), hoping that the values of V^* will closely match the values of V. Effect modification between V and the makers of $\mathcal{M}$ formally explains what we mean by "a better measurement of V".

Figure 14. Choosing a measurement method



Two remarks

Modified effects serve us well in various circumstances: choosing treatment, nullifying causal paths that are not of immediate interest, deconfounding, preventing some colliding bias, and reducing information bias. It is difficult to find another feature of causal reality that offers so much, except causality itself.

In most of the examples the modified effect was null for some value of the modifier and non-null otherwise. Again, the dichotomy between null and not null simplifies methodological discussions, but may be relaxed. It makes little difference, for instance, whether a confounding path is precisely nullified or is sufficiently weakened by randomization. Likewise, it is not that important whether the causal path from the offered treatment to the patient expectation is precisely nullified by blinding or only nearly so. In an indeterministic world, the pair "to modify" and "to not modify"—just like the pair "to cause" and "to not cause"—is a simple-minded descriptor of a continuous phenomenon. Whether referring to effect or to effect modification, the precise null is no more than another point on a continuum—hardly worthy of the attention it gets (commentary here).

References

1. Shahar E, Shahar DJ. Causal diagrams and three pairs of biases. In: *Epidemiology – Current Perspectives on Research and Practice* (Lunet N, Editor). www.intechopen.com/books/epidemiology-current-perspectives-on-research-and-practice, 2012:pp 31-62
2. Shahar E, Shahar DJ. On the definition of effect modification. *Epidemiology* 2010;21:587
3. Shahar E, Shahar DJ: Marginal structural models: much ado about (almost) nothing. *Journal of Evaluation in Clinical Practice* 2013;19:214-22

4. Shahar E, Shahar DJ: Causal diagrams and change variables. *Journal of Evaluation in Clinical Practice* 2012;18:143–148
5. Shahar E, Shahar DJ: On the causal structure of information bias and confounding bias in randomized trials. *Journal of Evaluation in Clinical Practice* 2009;15:1214-6
6. Shahar E, Shahar DJ: Causal diagrams, information bias, and thought bias. *Pragmatic and Observational Research* 2010;1:33–47