Chapter 5

Effect Modification

Sources of confusion

Effect modification is one of the trickier concepts in epidemiology for at least three reasons. First, it goes by another name—interaction—and the synonym portrays a different causal image. Saying that exposure to asbestos interacts with smoking to cause lung cancer sounds very different from saying that exposure to asbestos modifies the effect of smoking on lung cancer. Second, to explore effect modification, two routes may be taken: In one we ask whether the effect of a causal contrast depends on other causal assignments, whereas in the other we ask whether the joint effect of two causes (or more) conflicts with our expectation of what that joint effect might be. Yet as we'll realize later, these seemingly different routes are separated by only one line of algebra. As far as the math is concerned, they are not different at all. Third, we'll see that unless you declare a preference for one measurement scale of effects—ratios or differences—effect modification is almost always present. With few mathematical exceptions, if you don't find it on one scale, you are sure to find it on the other. Indeed, some writers argue that effect modification should be called effect-*measure* modification, linking the idea to the arbitrary scale on which effects are measured rather than to underlying causal reality. (Not believing in indeterministic causal forces, they have no burning commitment to ratio measures of effect.) Other writers argue that only the additive scale enriches our knowledge about causal pathways.

Time and again, the confusion and dispute are retraceable to the conflict between models of causation. In previous chapters we saw that determinism and indeterminism proved to be parallel routes of causal inquiry, disagreeing about the way causation works, about the definition of causal parameters, and about the purpose of randomization. In this chapter, we'll see how their disagreement is carried forward to effect modification.

One simple example and several complex thoughts

Before systematically exploring the mathematical trail, let's start at the end and show what effect modification means in a deterministic world. We'll follow again the triad of streptokinase, placebo and death, use the simplest possible example, and ask what may modify the effect of that drug on death. What makes the causal parameter take on different values?

Suppose that streptokinase never prevents the death of a person who suffered an ischemic stroke and that our interest centers on the proportion difference as the causal parameter—two assumptions that should not raise theoretical objections in too many deterministic minds. Recalling equation 4–3 from chapter 4, the causal parameter takes a simple form. In any group of patients, the effect of streptokinase on death is $P_{\textit{streptokinase causative}}$, the proportion of patients who belong to that deterministic class. Here is the proof:

Proportion Difference $_{\text{CAUSAL}} = P_{\textit{streptokinase causative}} - P_{\textit{streptokinase preventive}} =$

$P_{\textit{streptokinase causative}} - 0 = P_{\textit{streptokinase causative}}$

Now, to ask what may modify the effect of streptokinase on death is equivalent to asking what determines the proportion of stroke victims who belong to the class called *streptokinase causative*. The general answer is known, however, on two levels: metaphysical and causal. On a metaphysical level, Nature decides who belongs to that class and who does not, and she has no reason to fix that proportion according to the groups we assemble. $P_{\textit{streptokinase causative}}$ may take the value of 0.2 among 123 women, 0.4 among 456 men, 1.0 among 53 women in hospital A, and zero among 61 men in hospital B. Moreover, the various descriptors that follow the counts (men, women, hospital) may have nothing to do with the cited proportion.

On a causal level, the explanation lie in those component causes that together with streptokinase make up an unknown number of sufficient causes of death, perhaps millions of them. The value of $P_{\textit{streptokinase causative}}$ is determined, or modified, by any companion of streptokinase: Eliminate one contributor to one sufficient cause and you have one fewer death and a different proportion; add two contributors to complete two more sufficient causes and you have added two deaths, changing the proportion again. In a deterministic world the list of effect modifiers is as long as the list of component causes that conspire with streptokinase to bring about death, and their number may be as large as the number of stroke victims who are called *streptokinase causative*.

This is the place where extreme determinism meets stochastic causation, an extreme version of indeterminism (chapter 4). Although diametrically opposed, these two causal structures have two things in common. Neither is accessible to empirical science and both tailor the causal law to each victim of stroke who would die after receiving streptokinase. It is not the same law, however: In a stochastic world the causal law is the *propensity ratio* of death for the *i*-th patient (which we labeled C_i), whereas in a deterministic world it is the *names* of component causes in a sufficient cause of death of the *i*-th patient. In either case the true effect modifier and the "reason" for effect modification is the name of the patient. Effect modification, or what we previously called heterogeneity of effect, should be present everywhere—it is the governing rule of both structures.

Streptokinase, placebo, and the size of a stroke

After intense theoretical thinking, let's work out the algebra, which is fairly simple.

Let P denotes a measure of frequency of death (proportion, odds, or rate) in the streptokinase trial and just keep in mind that the answer to the question of what P estimates depends on the model of causation: In determinism, P estimates the frequency of patients who belong to some deterministic classes, whereas in indeterminism P estimates some causal propensities. Whatever their underlying reality may be, all four P's in Figure 5–1 should be known in the streptokinase trial provided that the size of the stroke was recorded. For example, P_3 may be the rate of death among placebo recipients who had suffered a large stroke.

		Treatment	
		Placebo	Streptokinase
Stroke size	Not large	P_1	P_2
	Large	P_3	P_4

Figure 5–1. Frequency of death (proportion, odds, or rate) in a trial of streptokinase, by treatment group and stroke size

Does the size of the stroke modify the effect of streptokinase on death?

To answer this question, we should estimate the effect of the causal contrast between streptokinase and placebo in two strata: patients who suffered a large stroke (Figure 5–1, second row) and patients who did not (first row.) Next, we'll have to compare the two effects and see whether they differ at all, or realistically speaking, by how much they differ. One minor question, however, must be answered first: Should we estimate these effects on a multiplicative scale (as ratios) or on an additive scale (as differences)? As I stated earlier and will show at the end, if the two effects happen to be similar on one measurement scale, they will almost always differ on the other. Algebra dictates diverging results regardless of which causal reality has generated the four frequencies of death.

Effect modification on an additive scale

If we decide to measure effects on an additive scale, then:
1) When the stroke is large, the effect of streptokinase as compared with placebo is P_4–P_3.
2) When the stroke is not large, the effect of streptokinase as compared with placebo is P_2–P_1.

And the question is therefore: Is P_4–P_3 different from P_2–P_1? (Or how large is the difference?)

If $P_4–P_3 = P_2–P_1$, or approximately so, we'll say that the effect on death of the causal contrast between streptokinase and placebo is *homogenous* across the two strata of stroke size *when that effect is measured on an additive scale*. Or stated differently: On an additive scale, the variable "stroke size" in its binary form is not an effect modifier of the effect of interest.

It may be interesting to recall the simple question we asked and compare it to the convoluted answer we ended up providing. We asked whether the size of the stroke modifies the effect of streptokinase on death, which sounded fairly straightforward. But to answer this question on solid footing we had to specify that: 1) "the effect of streptokinase" means the causal contrast between streptokinase and placebo; 2) effect was measured on an additive scale; 3) the candidate for the title "effect modifier" took a particular binary form. All three qualifications were essential because without them no single answer may be given. Change one qualification and you might get a different answer. Delete one qualification and you might have several different answers.

Back to the algebra. If however $P_4–P_3 \neq P_2–P_1$ and the difference is "large enough", we'll say that the size of the stroke modifies the effect on death of the causal contrast between streptokinase and placebo—again, when that effect is measured on an additive scale. If $P_4–P_3 > P_2–P_1 > 0$, the effect is larger (more harmful) when the stroke is large, and if $0 < P_4–P_3 < P_2–P_1$ the effect is smaller (less harmful) when the stroke is large. Notice two points, though: First, I described the effect as "larger" or "smaller" and not as "stronger" or "weaker" because a reference to causal strength requires in my mind a ratio scale of comparison. Second, these two stratum-specific effects may differ qualitatively—not only quantitatively. For example: if $P_4–P_3 > 0$ and $P_2–P_1 < 0$, streptokinase is actually helpful as compared with placebo when the stroke is not large, and if $P_4–P_3 > 0$ and $P_2–P_1 = 0$ streptokinase is not different from placebo when the stroke is not large.

Additive scale: reciprocity

It may have been natural to ask whether the effect on death of the causal contrast between streptokinase and placebo is homogenous across the two strata of stroke size. But it is also legitimate to turn the question upside down and ask whether the effect of the causal contrast between suffering a large stroke and suffering a smaller stroke is homogenous across the two treatment strata. A little algebra will show, however, that both questions must have the same answer, whatever that answer may be.

Suppose that the size of the stroke does *not* modify the treatment effect. That is,
$P_4 – P_3 = P_2 – P_1$.

Then, after adding P_3 to both sides and subtracting P_2 from both sides, we get
$P_4 – P_2 = P_3 – P_1$

Referring back to Figure 5–1, we can find out what the last equality says. It says that the effect of the stroke size when streptokinase is received ($P_4 – P_2$) is identical to the effect of the stroke size when placebo is received ($P_3 – P_1$). Or in other words, the variable "treatment group" does not modify the effect of the stroke size on death. If we next replace each equality sign with an inequality sign, we will conclude in parallel that effect modification by the stroke size implies reciprocal effect modification by the treatment group. If variable A modifies the effect of variable B on Z, then variable B modifies the effect of variable A on Z. (And it does not matter at all which causal assignments were randomly assigned, if any.)

It is common to forget the reciprocal property of effect modification and label one variable *exposure* and the other variable *effect modifier* as if these labels describe distinct causal roles. The algebra shows, however, that the labels may be switched without blinking an eye. But the key conclusion from the symmetry of effect modification is this: an effect modifier must be, in the first place, a causal variable! If it modifies the effect of another cause, it must also cause that effect.

Which has two interesting implications. First, we have just found another reason to argue that variables such as sex, race, and age are entitled to be called causal variables. Since nobody dispute their right to be called effect modifiers of a causal variable, they should qualify for the latter label as well. Second, when searching for modifiers of some cause-and-effect relation, we should consider only candidates that are thought to be causal variables themselves. And only after estimating their effects do we have a rationale for diving deeper to search for heterogeneity of effects. This idea explains, for example, why biomedical researchers often look for effect modification by age and rarely look for effect modification by blood type. Unlike age, blood type has not turned out (yet) to have many strong biological effects.

It may be easy to remember the following simple rule: For a given effect, only a causal variable can modify the effect of a causal variable.

Additive scale: an alternative route

Suppose that the size of the stroke and the treatment group are *not* reciprocal effect modifiers in the streptokinase trial (on an additive scale), which means that the following equality holds:

$P_4 - P_3 \quad = \quad P_2 - P_1.$

If so, $P_4 \quad = \quad P_3 + P_2 - P_1$

And, $(P_4 - P_1) \quad = \quad (P_3 - P_1) + (P_2 - P_1)$

Referring to Figure 5–1, we can decode the last equality according to the following causal language:

P_1: frequency of death for causal assignment	*placebo*	and	*not large stroke*
P_2: frequency of death for causal assignment	*streptokinase*	and	*not large stroke*
P_3: frequency of death for causal assignment	*placebo*	and	*large stroke*
P_4: frequency of death for causal assignment	*streptokinase*	and	*large stroke*

Therefore, each parenthetical term in equation xx-xx estimates the effect of a causal contrast as described below:

P_4–P_1: the contrast between the causal assignment behind P_4 (*streptokinase* and *large stroke*) and the causal assignment behind P_1 (*placebo* and *not large stroke*)

P_3–P_1: the contrast between the causal assignment behind P_3 (*placebo* and *large stroke*) and the causal assignment behind P_1 (*placebo* and *not large stroke*)

$P_2–P_1$: the contrast between the causal assignment behind P_2 (*streptokinase* and *not large stroke*) and the causal assignment behind P_1 (*placebo* and *not large stroke*)

There are many causal branches here, but they have a common trunk: P_1. All three causal contrasts share the same reference causal assignment, which is the left upper cell of Figure 5–1. Keeping this reference assignment in mind, we can translate equation xx-xx into the following language: The joint effect of receiving streptokinase and suffering a large stroke is identical to the sum of two effects: that of suffering a large stroke (when placebo is received) and that of receiving streptokinase (when the stroke is not large.)

If on the other hand the equality does *not* hold, then

either $(P_4 – P_1) > (P_3 – P_1) + (P_2 – P_1)$
or $(P_4 – P_1) < (P_3 – P_1) + (P_2 – P_1)$

which means that the joint effect of receiving streptokinase and suffering a large stroke is greater or smaller, respectively, than the sum of their separate effects.

The language I have just used is deceptively simple. I am on the verge of creating for you the image of two causes, streptokinase and a large stroke, that help each other on their path to cause death (the first inequality) or perhaps interfere with each other's path (the second inequality.) I am creating the illusion of two types of interaction between causes. One is *synergistic interaction* where two causes operate together, amplifying each other and accounting for more deaths than expected, and the other is *antagonistic interaction* where two causes interfere with each other (or perhaps overlap), accounting for fewer deaths than expected.

But the image is false. To create it in your mind, I conveniently picked two of four causal assignments, called them causes, and quietly dismissed their contrasts as an inferior species of causes. If on the other hand I had picked *receiving placebo* and *not suffering a large stroke* and called *them* causes, nobody would have fallen in the trap. Think for example about the statement "The joint effect of receiving placebo and suffering a stroke that is not large is greater than the sum of their separate effects." Where did the image of interacting causes go? And remember what we mean here by "the effect of receiving placebo": placebo versus streptokinase!

We like the image of interaction between causes because it rings like determinism, a causal model we may be programmed to prefer, and it's easy to bring up the image if we choose the "right" causal assignments—streptokinase and large stroke, for example. We can visualize how component causes join hands—interact—to make up one sufficient cause of an event. The word interaction indeed resonates well in the deterministic mind.

Sparing you the math, the truth is this. If the world is deterministic and, say, $(P_4–P_1) > (P_3–P_1) + (P_2–P_1)$, it is possible to deduce under certain assumptions that streptokinase and suffering a large stroke indeed joined hands in some sufficient causes of death. For this reason, perhaps, deterministic writers often argue that interaction should be measured and detected on an additive scale. They say that interaction on that scale, unlike its counterpart on a multiplicative scale, tells us about biological reality—conveniently forgetting that determinism has restricted all causal inference to a finite population. Unless "biological reality" means the exclusive biology of an arbitrary target population, they cannot have it both ways. Either inference on biological interaction is shredded when the target population expires (just like the fate of its deterministic causal parameter), or they have to reconcile determinism with causal knowledge that is not restricted to a target population. Or better: choose indeterminism where there are causal propensities and no target populations.

Effect modification on a multiplicative scale

Much of the algebra of the additive scale can be translated into the multiplicative scale by replacing every summation with multiplication, every subtraction with division, and every "additive" with "multiplicative." To avoid repetition over several parallel sections, the story is summarized in one. But it is not less important.

When the stroke is large, the effect of streptokinase (versus placebo) is P_4 / P_3.
When the stroke is not large stroke, the effect of streptokinase (versus placebo) is P_2 / P_1.

And the question is therefore: Is P_4 / P_3 different from P_2 / P_1? (Or how large is the difference?)

If $P_4 / P_3 = P_2 / P_1$, or approximately so, we'll say that the effect on death of the causal contrast between streptokinase and placebo is *homogenous* across the two strata of stroke size *when that effect is measured on a multiplicative scale*. And if P_4 / P_3 is "sufficiently different" from P_2 / P_1, we will claim heterogeneity of effect and call the size of the stroke an effect modifier. Following minor algebraic rearrangement, we can derive again the reciprocal property of effect modification (P_4/P_2 is equal to P_3/P_1 or they are "sufficiently different"), thereby treating the stroke size as the effect of interest and the treatment group as a possible effect modifier.

To compare the strength of two effects, we stay on the ratio scale and divide one effect estimate by the other. For example, the ratio of P_4/P_2 to P_3/P_1 will tell us how much stronger (or weaker) the effect of the stroke size is when streptokinase is received than when placebo is received.

Regardless of whether effects are measured as ratios or differences, you may find it helpful to visualize the reciprocal approaches to effect modification by splitting the table of Figure 5–1 in two ways (Figure 5–2): In the left panel we visualize the treatment effect in two strata of stroke size whereas in the right panel we visualize the effect of the stroke size in two treatment strata.

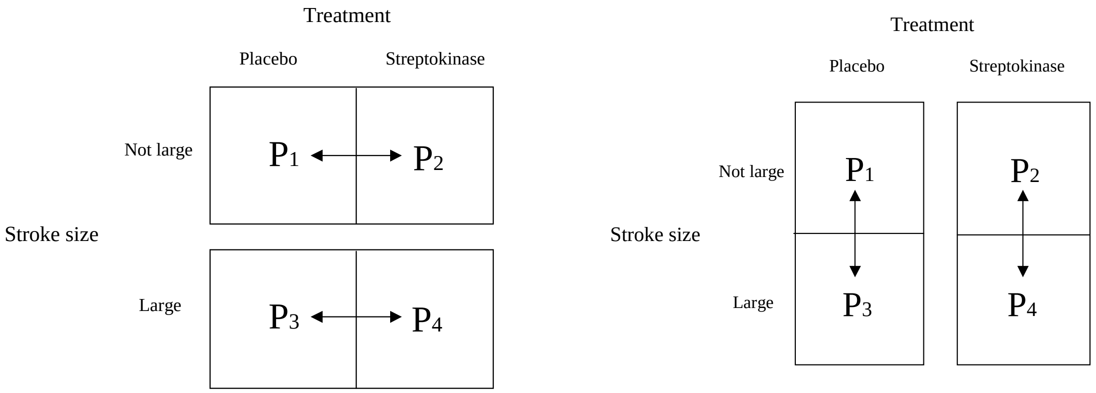

Figure 5–2. Visualizing the reciprocal approaches to effect modification

The route to interaction on a multiplicative scale parallels its counterpart on an additive scale.

Interaction is absent if

$P_4 / P_3 = P_2 / P_1$ (or, alternatively, $P_4 / P_2 = P_3 / P_1$)
Then, $P_4 = P_3 \text{ x } (P_2 / P_1)$
And, $(P_4 / P_1) = (P_3 / P_1) \text{ x } (P_2 / P_1)$

The last equality says that the joint effect of receiving streptokinase and suffering a large stroke is identical to the product of two effects: that of a large stroke (when placebo is received) and that of receiving streptokinase (when the stroke is not large.) Again, if the equality sign changes to an inequality sign, we'll conclude that the joint effect is either greater than we might have expected (synergism) or smaller than we might have expected (antagonism).

The multiplicative version of interaction does not seem to resonate as well as the additive version, maybe because asking two causes to produce more than their product sounds like asking too much. Psychology aside, we don't really need to worry here about the interpretation of any algebraic conversion. The idea of effect modification of the propensity ratio and the search for effect modifiers naturally follow the theoretical foundation for indeterministic causal parameters (chapter 4).

Numerical examples

Let's illustrate the algebra of effect modification and interaction by hypothetical examples.

		Treatment	
		Placebo	Streptokinase
Stroke size	Not large	10 (P_1)	40 (P_2)
	Large	20 (P_3)	80 (P_4)

Figure 5–3. Hypothetical death rate (per 100 person-years at risk) in a trial of streptokinase, by treatment group and stroke size

Does the size of the stroke modify the effect of the treatment group?

On an additive scale, it does: 80–20 ≠ 40–10. The effect of the treatment group is larger (here, unfavorable to streptokinase as compared with placebo) when the stroke is large than when it is not. Or, in reciprocal language, treatment group modifies the effect of the stroke size: 80–40 ≠ 20–10: the effect of the stroke size is larger (unfavorable to large stroke as compared with a smaller stroke) when streptokinase is received than when placebo is received.

On a multiplicative scale, however, there is no effect modification. The effect of the treatment group is as strong when the stroke is large as when it is not: 80/20 = 40/10. Likewise, and as is expected from the reciprocal property, the treatment group does not modify the effect of the stroke size: 80/40 = 20/10.

Switching to the language of interaction, the question will be reworded as follows: Do *streptokinase* and *large stroke* interact to cause death?

On an additive scale they do—synergistically: (80–10) > (20–10) + (40–10)
On a multiplicative scale, they don't: (80/10) = (20/10) x (40/10)

It is possible to come up with a few examples where the two scales produce a similar result (no interaction on either, for example) but in most examples they will not. If effect modification (or interaction) is not found on one scale, it will be found on the other. And when it is found on both, it is not uncommon to find synergistic interaction on the additive scale that translates to antagonistic interaction on the multiplicative scale. That is,

$(P_4 - P_1) \; > \; (P_3 - P_1) + (P_2 - P_1)$
yet
$(P_4 / P_1) \; < \; (P_3 / P_1) \text{ x } (P_2 / P_1)$

For example: if P_1=10, P_2=30, P_3=40, and P_4=80 (Figure 5–4) the interaction is synergistic on the additive scale, 80–10 > (30–10) + (40–10), but antagonistic on the multiplicative scale: 80/10 < (30/10) x (40/10).

		Treatment	
		Placebo	Streptokinase
Stroke size	Not large	10 (P_1)	30 (P_2)
	Large	40 (P_3)	80 (P_4)

Figure 5–4. Hypothetical death rate (per 100 person-years at risk) in a trial of streptokinase, by treatment group and stroke size

It is interesting to see what the last example mean in the language and math of effect modification. If we consider the size of the stroke as the effect modifier, then

on the additive scale	80–40 > 30–10
whereas on the multiplicative scale	80/40 < 30/10

Looking at Figure 5–4, we'll decode these inequalities as follows: On the additive scale, the effect of the contrast between streptokinase and placebo is *larger* when the stroke is large than when it is not (40 versus 20 extra deaths per 100 person-years). On the multiplicative scale, however, the effect of the same contrast is *weaker* when the stroke is large than when it is not (a propensity ratio of 2 versus 3). So—you are probably asking—who cares that the causal propensity is weaker if more people die? How can the term "propensity ratio" compete with the words "extra deaths"? The former is hypothetical, theoretical, unreal. The latter words sound like reality—dead people.

What you have just read or independently thought is no more than another example of the unbridgeable gap between determinism and indeterminism. And these kinds of examples undoubtedly play into the hands of determinism, generating sympathy for its trail and a dismissive attitude toward the other trail. What you should not forget, however, is that the human inclination to prefer one computation to another does not alter causal reality. If causal propensities make up the world's structure and we want to discover it, we should fit our numerical tools to that structure and not settle for tools that fit our compassionate mind. But there is more to remember. Besides the idea of extra deaths, the additive scale and determinism carry along several consequences that are not so appealing to the human mind (chapter 4): a proportion difference of zero when streptokinase kills 10% of the patients and saves another 10%; an arbitrary target population in which those extra deaths occur; and no biological inference when the target population is no longer alive. Keeping these facts in the background might help us to look at the last two examples through different lenses.

Beyond the binary world

When the effect makes up a continuous variable, P_1, P_2, P_3, and P_4 in Table 5–1 are no longer frequencies. They turn into means—either arithmetic means or geometric means. For example, P may be the mean of a variable that measures muscle strength on a scale from 0 to 100. Remember, however, that if you chose the arithmetic mean you are committed to looking for effect modification on the additive scale (of the mean difference) and if you chose the geometric mean, you are committed to looking for effect modification on the multiplicative scale (of the geometric mean ratio). Examples of both computations will be shown in chapters 8 and 9.

Things get more complicated when one of the two causal variables, or both variables, contain more than two categories or when they are continuous variables themselves. We'll address these complex situations in later chapters.

Effect modification: the indeterministic viewpoint

Recalling chapter 4, indeterminism prescribes three possible structures for the effect of a causal contrast: an identical propensity ratio across all background propensities (homogeneity of the causal parameter), varying effect by any number of background propensities, and varying effect by personal identity (stochastic causation.) For me, the empirical expression of effect modification (say, $P_4/P_3 > P_2/P_1$) temporarily supports the second of these three structures: the effect of one causal contrast depends on the background causal propensity that was generated by other causal assignments. In an indeterministic world, our search for effect modification and for effect modifiers does not follow an optional side trail of causal inquiry. It is part of the main road and the only available method by which we may revise our fallible claim that the first structure holds—that one number says it all about one causal contrast and its effect.

Being indeterminist, I am neither indifferent to the measurement scale nor am willing to concede to those who argue for the additive scale. Effect modification, just like unmodified effects, should be unapologetically explored on the multiplicative scale—as ratios—because that's the scale on which we can unambiguously compare the strength of causal forces. And for those of us who are not persuaded by the argument, statistics has already decided the matter on a technical point. As you will see later, most commonly used regressions models with which we estimate effects and heterogeneity of effects are inherently multiplicative, forcing us to estimate ratios rather than differences: logistic regression (odds ratio), Poisson regression (rate ratio), and Cox regression (hazard ratio).