12. Multicausality: Effect Modification

Issues in characterizing the combined effect of two or more causes of a disease (or, equivalently, the effect of one factor in the presence or absence of other factors).

Multicausality

The rise of the germ theory of disease brought with it the paradigm of specificity of disease causation, in which diseases were specific entities and each specific disease had a specific cause. Since identifiable microorganisms could be linked to specific clinical syndromes and natural histories, this paradigm contributed to the dramatic progress in medical microbiology and development of antibiotics, which have transformed human vulnerability to infectious disease. The doctrine of specific causation proved something of a hindrance, however, in the study of noninfectious disease, notably in appreciating the health effects of tobacco smoke.

Now that the concept of multifactorial disease is fully accepted, we should perhaps adopt a more relativist perspective, in which specificity of causation varies according to the "disease" (used hereinafter to refer to any outcome of interest) and its definition, the type of causal agents or factors we wish to consider, and on the stage in the causal process. John Cassel invited this way of thinking when he described tuberculosis, a hallmark of the revolution in bacteriology, as a multifactorial disease in regard to various characteristics of the host and his/her social environment. The resurgence of mycobacterial tuberculosis in the United States during the 1980's, as a result of such factors as the spread of HIV, the rise in homelessness, and the reduction of funding for tuberculosis control, illustrates the importance of host and environmental factors for this disease.

The upsurge in syphilis in the southeastern region of the U.S. during a similar period provides another example. In the chapter on Phenomenon of Disease, syphilis served as an example of a disease defined by causal criteria and thus definitionally linked to a specific microorganism, *Treponema pallidum*. Syphilis infection can induce a great variety of symptoms and signs, so great that it has been called the "Great Imitator" (by Sir William Osler, I believe, who I think also wrote "Know syphilis and you will know all diseases"). Given the diversity of ways in which syphilis manifests, it is fortunate that we do not need to rely on manifestational criteria to define syphilis. Nevertheless, although defined in relation to its "cause", syphilis can also be considered a multifactorial disease, since risk of syphilis is related to personal and contextual factors such as number and type of sexual partners, use of condoms, exchange of sex for drugs or money, use of crack cocaine, access to care, proficiency of clinicians, effectiveness of public health services, degree of social stigma, racism, and limited resources devoted to developing a vaccine. Since syphilis is not transmitted in every unprotected exposure, there may be transmission and immune factors to add to this list.

Similarly, coronary heart disease is a classic multifactorial disease, with an ever-growing list of risk factors that includes at least atherogenic blood lipid profile, cigarette smoke, elevated blood pressure, sedentary lifestyle, diabetes mellitus, elevated plasma homocysteine, and insufficient intake

of dietary antioxidants. However, coronary artery disease is a clinically-defined entity that develops from a composite of changes in the coronary arteries. As our understanding of the pathophysiology and pathogenesis of coronary heart disease becomes more refined, researchers may eventually decide that it is more useful to subdivide this complex disease entity into its specific pathogenetic processes, which include certain types of injury to the coronary endothelium, growth of atheromas, and thrombus formation. These different pathologies could be defined as separate diseases, even though clinical manifestations usually require more than one to be present.

The one-variable-at-a-time perspective

Epidemiologists, however, typically focus on a single putative risk factor at a time and only sometimes have the opportunity to focus on specific pathogenetic processes. One reason for this is that epidemiology is in the front lines of disease control, and it is often possible to control disease with only a very partial understanding of its pathophysiology and etiology. Once it was demonstrated that cigarette smoking increased the risk of various severe diseases, including lung cancer, coronary heart disease, and obstructive lung diseases, many cases could be prevented by reducing the prevalence of smoking even though the pathophysiologic mechanisms were largely unknown. Once it was found that AIDS was in all probability caused by an infectious agent and that unprotected anal intercourse greatly facilitated its transmission, effective preventive measures could be taken even before the virus itself was identified and the details of its pathogenicity unravelled.

Thus, epidemiologists often find ourselves taking a "one-variable-at-a-time" approach to diseases of unknown and/or multifactorial etiology. Lacking the knowledge needed to work from a comprehensive model of the pathophysiologic process, epidemiologists attempt to isolate the effects of a single putative risk factor from the known, suspected, or potential effects of other factors. Thus, in the preceding chapter we examined how the effects of one factor can be misattributed to another factor ("guilt by association") and considered ways to control for or "hold constant" the effects of other risk factors so that we might attribute an observed effect to the exposure variable under investigation.

Another consequence of the one-variable-at-a-time approach is the phenomenon that an association we observe may vary according to the presence of other factors. From our ready acceptance of multicausation, we have little difficulty entertaining the idea that some disease processes involve the simultaneous or sequential action of more than one factor or the absence of a preventive factor. Indeed, with the growth of genetic knowledge all disease is coming to be regarded as a product of the interaction of genetic and environmental (i.e., nongenetic) factors.

But from the one-variable-at-a-time perspective, our window into these interdependencies comes largely from measures of association and impact for each particular risk factor-disease relationship. Thus, if two factors often act in concert to cause disease, we will observe the risk difference for one of the factors to differ depending upon the level of the other factor. It may therefore be important to control for factors that may modify a measure of effect of the exposure of primary interest. Control may be necessary even if the susceptibility factor cannot itself cause the disease and so would not qualify as a potential confounder.

Interdependent effects

The preceding chapters have largely dealt with situations involving a single exposure and a single outcome. The chapter on standardization of rates and ratios and the chapter on confounding concerned the need to control for a variable, such as age or a second exposure, so that comparisons could focus on the exposure of primary interest. We referred to the interfering variable as a confounder or potential confounder – essentially a nuisance variable – that threatened to interfere with our investigation of the primary relationship of interest.

We now want to consider another role for a second exposure variable. That role is involvement in the pathophysiologic process or in detection of the outcome in concert with or in opposition to the study factor (an exposure of primary interest). One of the factors may be regarded as a co-factor, a susceptibility factor, a preventive factor, or something else whose effect is entwined with that of the study factor.

Confounding, as we saw in the preceding chapter, results from an association between the exposure and the confounder. But the effects of these two exposures on the disease can be independent of one another. In fact, in the (hypothetical) Type A example, the exposure had no effect at all. In this chapter we are interested in exposures whose effects on the outcome are interdependent.

There are innumerable scenarios we can think of where such interdependence occurs. One entire category of interdependence involves genetic diseases whose expression requires an environmental exposure. For example, favism is a type of anemia that is caused by consumption of fava beans in people with reduced glucose-6-phosphate dehydrogenase (GPDH) activity. The anemia develops only in response to a constituent of fava beans, but people with normal GPDH activity are unaffected.

Another category of interdependence is that between exposure to infectious agents and immune status. Measles occurs only in people who have not already had the disease and rarely in people who have received the vaccine. People whose immune systems have been weakened by malnutrition or disease are more susceptible to various infectious agents, and HIV infection can render people vulnerable to a variety of infections called "opportunistic" because they occur only in immunocompromised hosts.

Causal chains that involve behaviors provide many illustrations of interdependency in relation to outcomes. Condoms reduce STD risk only when the sexual partner is infected. Airbags provide lifesaving protection to adult-size passengers involved in frontal crashes but can harm small passengers and provide less protection to persons not wearing a lap belt. Handguns are probably more hazardous when in the possession of people with poor anger management skills.

Since very few exposures cause disease entirely by themselves (rabies virus comes close), nearly every causal factor must modify the effect of other causal factors and have its effect modified by them. When these other factors are unidentified, they are generally regarded as part of the background environment, assumed to be uniformly distributed, and hence disregarded. Part of the challenge of

epidemiologic research is to identify major modifying factors that are not uniformly distributed, so that differences in findings across studies can be understood.

The terminology thicket

Even more than other areas of epidemiology, learning about how epidemiologists approach interdependent effects is complicated by a two decades old controversy about definitional, conceptual, and statistical issues and by a terminology that is as heterogeneous as the enrollment in a large class in introductory epidemiology! The terms epidemiologists have used to discuss interdependent or "joint" effects include: "synergy", "synergism", "antagonism", "interaction", "effect modification" (and "effect modifier"), and most recently "effect measure modification".

"**Synergy**" or "**synergism**" is the term applied to a situation in which the combined effect of two (or more) factors is materially greater than what we would expect from the effect of each factor acting alone. "**Antagonism**" refers to the reverse situation, where the joint effect is materially less than what we would expect. Synergism and antagonism are both types of "interaction".

The factors involved in an interdependent relationship can be regarded as having their effects modified by each other, which gives rise to the terms "effect modification" and "effect modifier". Sometimes the adjectives "quantitative" and "qualitative" are employed to distinguish between situations where the modifying variable changes the direction of the effect of the primary exposure or changes only the magnitude of effect. In **quantitative effect modification**, the modifier may strengthen or weaken the effect of the primary exposure, but the direction of effect does not change. In **qualitative effect modification**, the exposure either (1) increases risk in the presence of the modifier but reduces risk in its absence or (2) increases risk in the absence of the modifier but reduces risk in its presence. Although I first heard this distinction in a seminar presented by Doug Thompson, he more recently has referred to qualitative effect modification as a **crossover effect** (Thompson 1991).

Somewhere I picked up (or made up) the term "absolute effect modification" to refer to situations where the effect of at least one factor occurs only in the presence (or absence) of another factor. In such cases the first factor has no independent effect. In contrast, "relative effect modification" refers to situations where both factors have independent effects on risk regardless of the presence or absence of the other, but their joint effect is different from what one expects from their individual effects.

[Since more than two factors are generally involved, that means that, for example, variable A can be an absolute modifier of the effect of variable B (B has no effect without A) and a relative modifier of the effect of variable C (C has an effect without A, but its effect is stronger [weaker] in the presence of A). Whether B and/or C are absolute or relative modifiers of depends, in turn, on whether or not A has an (independent) effect on risk without B and/or C. But we are getting ahead of ourselves here.]

All of this terminology would be simply a matter of memorization were it not for one central difficulty. That difficulty arises in operationalizing the above concepts through the use of

epidemiologic data. Put simply, there is no simple connection between the concepts expressed above and the epidemiologic measures we have been using. Partly because of this disconnect, the terms "interaction" and "effect modification" have been employed with different meanings at different times by different authors (and sometimes by the same author). Thompson (1991:p221) says that the two terms have different "shades of meaning" but (wisely) uses the two terms interchangeably.

Previous editions of this chapter attempted to reduce terminology confusion by following the usage in the first edition of Rothman's text *Modern Epidemiology*. Rothman used the term "biological interaction" to refer to synergy or antagonism at the level of biological mechanisms, such as that in the favism example. He used the term "effect modification" to refer to data that give the appearance of joint effects that are stronger or weaker than expected (statistical interaction falls into this category). The second edition of *Modern Epidemiology* introduces a new term, "**effect measure modification**", with the purpose of reducing the tendency to link data and biology through the use of the same word. Kleinbaum, Kupper, and Morgenstern used the terms "**homogeneity**" and "**heterogeneity**" to indicate similarity or difference in a measure across two or more groups. These neutral terms, which carry no connotation of causation, may be the safest to use.

Statistical interaction

The term "interaction" has an established and specific meaning in statistics, where it is used to characterize a situation where effects are not additive. (Statisticians have the significant advantage of being able to use the term "effects" without a causal connotation.) For example, **analysis of variance** is used to compare the means of one variable (e.g., blood pressure, BP) between two or more populations. If we are concerned that BP is influenced by another variable (e.g., body mass index, BMI) and that the two populations have different BMI distributions, we may want to adjust the BP comparison for BMI. (The idea is similar to our computation of a standardized rate difference to compare mortality rates in two populations.) If the relationship between BP and BMI is linear, then the method of adjustment is called **analysis of covariance** and can be illustrated as two lines on a pair of axes (see left side of figure).

The vertical distance between the lines represents the **adjusted difference** in mean BP between the two populations. Unless the two lines are parallel, however, the distance between them will vary with the level of BMI. The lines will be parallel when the slope of the relationship between BP and BMI is the same in the two populations, i.e., the strength of the association between blood pressure and BMI is the same in the two populations.

When the two lines are parallel, the blood pressures in the two populations can be represented by an equation with three terms on the right-hand side – a constant (**a**), a variable (POP) indicating the population in which the relationship is being estimated, and BMI, e.g.,

$$BP = a + b_1\,POP + b_2\,BMI$$

in which a, b_1, and b_2 will be estimated through a procedure called "**linear regression**".

Since the indicator variable (POP) is usually coded as 0 for one population and 1 for the other, the equations representing blood pressures are:

$$BP = a + 0 + b_2\,BMI \qquad (POP=0)$$

in one population and:

$$BP = a + b_1 + b_2\,BMI \qquad (POP=1)$$

in the other.

b_1 is then the vertical distance between the two lines, which corresponds to the adjusted difference in mean blood pressure between the populations. b_2 is the slope of the relationship between BP and BMI, i.e. the number of units increase in BP associated with a one-unit increase in BMI. This term accomplishes the adjustment needed for BMI. "a" is a constant that is usually needed to move the lines to their correct vertical position.

In the right side of the figure, the two lines are not parallel – there is **interaction**. Since the distance between the lines varies according to the level of BMI, the distance cannot be stated as a single number. In the presence of interaction, the linear model for blood pressure requires the addition of an "**interaction term**" to represent the varying distance between the lines:

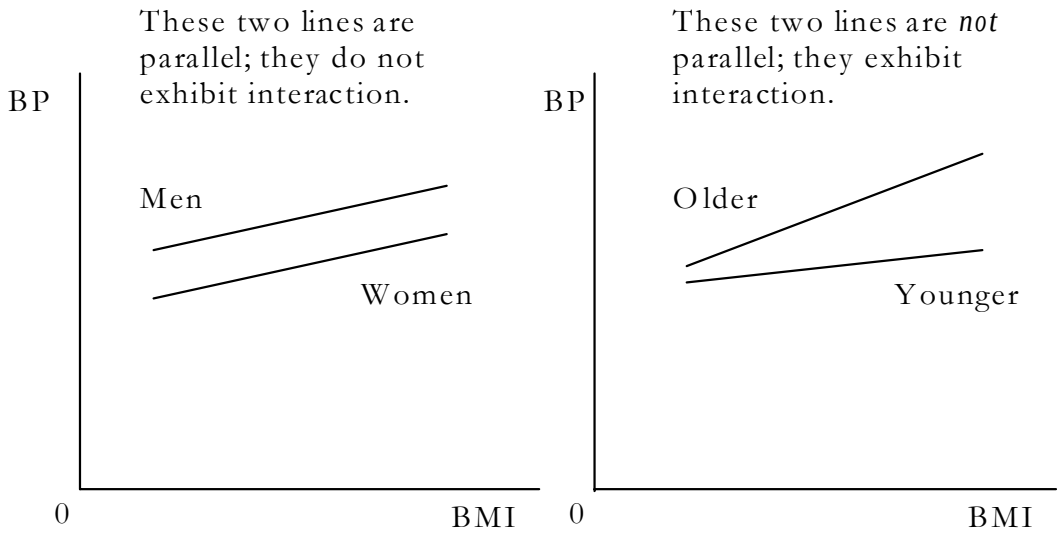
$$BP = a + b_1\,POP + b_2\,BMI + b_3\,(POP)\,(BMI)$$

With POP coded as 0 or 1, the first population will still have its blood pressures modeled by: $BP = a + b_2\,BMI$. However, the data in the second population will be modeled as:

$$BP = a + b_2 + b_2\,BMI + b_3\,BMI \qquad (POP=1)$$

b_3 represents a further adjustment to account for the lack of parallelism and thus the inability of b_1 alone to represent the distance between the lines. The difference between the two populations will be stated as $(b_1 + b_3\,BMI)$, so that it will be different for different levels of BMI.

Illustration of statistical interaction



If the figure on the left represents the relationship between blood pressure (BP) and body mass index (BMI) in men (upper line) and women (lower line), then the graph shows that the association of body mass and blood pressure is equally strong in both sexes – a one-unit increase in body mass index in men and a one-unit increase in women both are associated with the same increase in blood pressure. Therefore there is no (statistical) interaction.

In contrast, if the figure on the right represents the relationship in older people (upper line) and younger people (lower line), then the graph indicates an interaction between body mass index and age – a one-unit increase in body mass index in older people is associated with a larger increase in blood pressure than is a one-unit increase in younger people.

Statisticians use the "interaction" to refer to the latter situation, where the equations for different groups differ by a variable amount on a given scale (e.g., interaction may be present on the ordinary scale but not on the log scale).

Biological interaction

Epidemiologists are more interested in what Rothman and Greenland call "biological interaction". Biological interaction refers to interdependencies in causal pathways, such as those discussed at the beginning of this chapter. Such interdependencies – situations where one factor may potentiate or inhibit the effect of another – have implications for understanding of disease etiology or effectiveness of treatments or interventions. Laboratory researchers can readily observe such interdependencies, but epidemiologists must content ourselves with analyzing clinical or population data.

Over two decades ago (Causes, *Am J Epidemiol*, 1976), Rothman introduced a diagrammatic representation of multicausality in this biological or other mechanistic sense. He has continued to elaborate this schematic model and uses it to illustrate and explain relationships between epidemiologically-perceived relationships and "biological relationships".

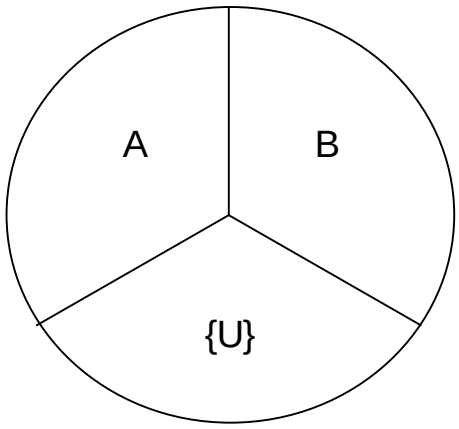
Rothman's model envisions causal pathways ("sufficient causes") as involving sets of "component causes". A "sufficient cause" is any set of component causes that simultaneously or sequentially bring about the disease outcome. "Component causes" are the individual conditions, characteristics, exposures, and other requisites (e.g., time) that activate the available causal pathways. Since there are always causal components that are unknown or not of interest for a particular discussion, sufficient causes include a component to represent them. Let us explore the way Rothman's model works.

"Cause" - (1) an event, condition or characteristic that plays an essential role in producing the occurrence of the disease (this is a "component cause"); or (2) a constellation of components that act in concert.

"Sufficient cause" - Set of "minimal" conditions and events that inevitably produce a disease; none of the conditions is superfluous; most of the components are unknown.

"Necessary cause" - A causal component that must be present for the disease to occur.

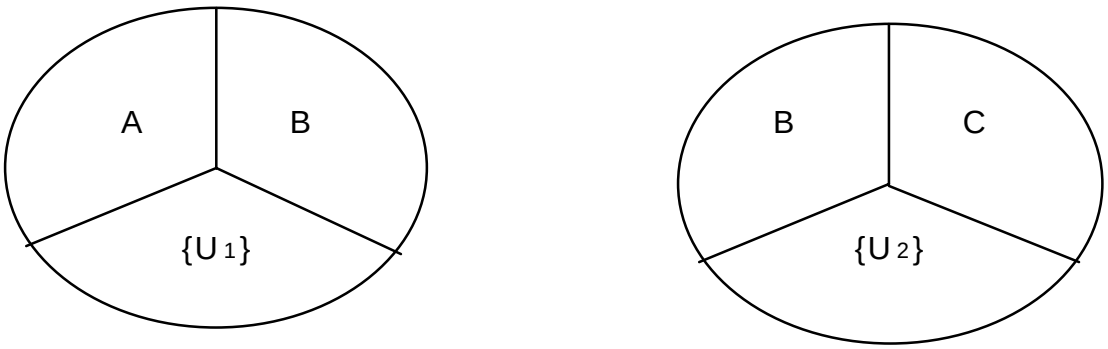
The circle below represents a sufficient cause, e.g., a pathway, chain, or mechanism that can cause a particular disease or other outcome. If all components are present, then the disease occurs (on analogy with the game Bingo). A and B represent component causes. For this sufficient cause to come into play, both A and B must be present. {U} represents the unknown background factors that also must be present for this sufficient cause to operate.



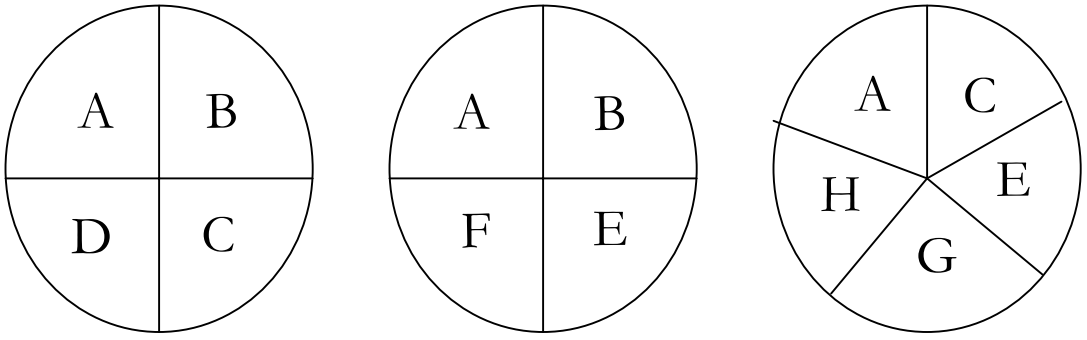
If this diagram (model of biological, chemical, physical, psychological, etc. reality) represents the primary or only pathway to the outcome, then component causes A and B have interdependent effects. Each component cause must be present for the other to have its effect. We could say that they are synergistic. The favism situation could be represented in this way, with A representing fava

bean intake and B representing genetically-determined reduced glucose-6-phosphate dehydrogenase activity. If this sufficient cause is the only causal pathway by which the disease can occur, then this synergism is absolute: without A, B has no effect; with A, B does if the remaining components {U} are present; without B, A has no effect; with B, A does (when {U} are present). (If either factor is preventive, then A or B represents its absence.)

If there were additional causal pathways containing B but not A, then the absence of A would not completely eliminate the effect of B. The latter situation, illustrated below, might be characterized as intermediate, partial, or relative synergism. B can now affect disease risk even in the absence of A.



A has thus become a relative modifier of the effect of B. B, however, remains an absolute modifier of the effect of A, because A has no effect in the absence of B. We may also note that component cause B is a necessary cause, since there is no sufficient cause (causal pathway) that can operate unless B is present.



In this diagram, G and H exhibit absolute synergy, since neither has an effect in the absence of the other. B and C exhibit partial synergy with respect to each other, since their combined effect exceeds what would be expected from knowledge of their separate effects.

Applying Rothman's model to epidemiologic concepts (induction period) and measures

In our discussion of natural history of disease, we defined the induction period as the time between initial exposure and the development of the disease. Since causal pathways involve multiple component causes, though, in Rothman's model the concept of induction period applies to component causes, rather than to the disease. The induction period in respect to a particular component cause is the time usually required for the remaining component causes to come into existence. If necessary, one component cause can be defined as a period of time (e.g., for a microbial pathogen to multiply). By definition, the induction period for the last component cause to act has length zero.

Another and even more fundamental issue is that in multicausal situations, disease occurrence, extent, association, and impact all depend upon the prevalence of the relevant component causes in the populations under study. While we have previously acknowledged that the incidence and/or prevalence of a disease or other phenomenon depends upon the characteristics of the population, we have not examined the implications of this aspect for other epidemiologic measures. For example, we have generally spoken of strength of association as though it were a characteristic of an exposure-disease relationship. But though often treated as such, strength of association is fundamentally affected by the prevalence of other required component causes, which almost always exist.

Rothman's model helps to illustrate these relationships in situations where biological interdependency (used as a general term to signify any causal interdependency) is present. A basic point is that disease incidence in persons truly unexposed to a study factor indicates the existence of at least one sufficient cause (causal pathway) that does not involve the study factor. If exposed persons have a higher prevalence of the component causes that constitute this sufficient cause, their disease rate will be higher. This process is the basis for confounding to occur.

Second, since very few exposures are powerful enough to cause disease completely on their own, the rate of disease in exposed persons will also depend upon the prevalence of the other component causes that share pathways (sufficient causes) with the exposure. Measures of association and impact will therefore also depend upon the prevalence of other component causes, since these measures are derived from incidence rates.

Third, if two causal components share a causal pathway, then the rarer of the two component causes will appear to be a stronger determinant of the outcome, especially if the remaining component causes are common. As in economics, the limiting factor in production experiences the strongest upward pressure on price.

Fourth, proportion of disease attributable to a component cause (i.e., its ARP) depends upon the prevalence of the other component causes that share the causal pathway(s) to which it contributes. This result is so because if the strength of association depends upon prevalences, then so must the ARP. However, the ARP's for the various component causes are not additive and will often sum to more than 1.0. For example, if two component causes are in the same causal pathway, then the entire risk or rate associated with that pathway can be attributed to each of the two components. The absence of either component prevents the occurrence of the outcome.

Phenylketonuria example

An example of these relationships, from the article referred to earlier (Causes, *Am J Epidemiol* 1976; 104:587-92), is the causation of phenylketonuria (PKU), a condition that, like favism, is linked to a dietary factor (phenylalanine, an amino acid) and a genetic defect. Infants with the PKU gene who ingest more than a minimal amount of phenylalanine develop serious neurologic effects including mental retardation. The "causal pie" for this example would be the same as the first one in this chapter, with A representing the PKU gene and B representing dietary phenylalanine.

Since Western diets typically contain phenylalanine, in the absence of specific preventive measures (universal screening of newborns and institution of a special diet) nearly all infants with the PKU gene develop clinical manifestations. The risk ratio for the PKU gene is therefore enormous; the PKU gene is a "strong" cause. In contrast, phenylalanine is a "weak" cause, since nearly all infants are exposed to it and only a tiny proportion develop clinical PKU. However, in a society in which a large proportion of the population have the PKU gene and infant diets rarely contain phenylalanine, then dietary phenylalanine will appear as the strong cause and the PKU gene as the weak cause! (Recall: "any measure in epidemiology is a weighted average . . .".).

Numerical example - favism

To explore these ideas further, let us construct a numerical example. Suppose that in a population of size 10,000,000, there are two sufficient causes of favism, one that involves both GPDH deficiency and fava bean intake, and a second that involves neither of these factors. Assume:

- 1% of the population (100,000 persons) have GPDH deficiency;
- 20% (2,000,000) of the population consume fava beans;
- These two factors are distributed independently of one another, so that 20,000 people have both factors (20,000 = 1% of the 2,000,000 fava bean = 20% of the 100,000 GPDH deficient persons).
- All remaining component causes {U} needed to lead to favism through the first sufficient cause are simultaneously present in 10% of persons, independent of their other risk factors;
- The sufficient cause that does not involve fava beans or GPDH deficiency occurs in 0.03% of the population, again independent of other factors/component causes. (We are assuming that the definition of favism does not require involvement of fava beans themselves.)

In this situation, the first sufficient cause will act in the expected 1% x 20% x 10% = 0.02% of the population in whom all these components are present, i.e., 2,000 cases. The second sufficient cause will operate in 3,000 persons, regardless of GPDH deficiency and/or fava beans. The table below shows what we can expect to observe in various subsets of the population.

Incidence of favism by population subgroup

Sub-population	N	Incidence	Cases
People who do not eat fava beans and do not have GPDH deficiency; [N = 80% x 99% x 10,000,000; cases come only from the 2nd pathway]	7,920,000	0.03%	2,376
People who eat fava beans but do not have GPDH deficiency [N = 20% x 99% x 10,000,000; cases come only from the 2nd pathway]	1,980,000	0.03%	594
People with GPDH deficiency who do not eat fava beans [N = 1% x 80% x 10,000,000; cases come only from the 2nd pathway]	80,000	0.03%	24
People with GPDH deficiency who eat fava beans N= 1% x 20% x 10,000,000; 10% (2,000 cases) occur in the 10% with the remaining component causes; also, 0.03% of the 20,000 (6 cases) get favism through the second pathway; (0.6 people would be expected to have both pathways acting so are subtracted from the above total)]	20,000	10.03%	2,005.4
Total	10,000,000	0.05%	4,999.6

From this table we can compute (crude) incidences and incidence ratios for each exposure:

Incidence and incidence ratios of favism (crude)

GPDH deficiency		
Present (2,030 cases / 100,000 people)	2.03%	
Absent (2,970 / 9,900,000 people)	0.03%	
Incidence ratio		67.67
Eat fava beans		
Yes (2,600 cases / 2,000,000)	0.13%	
No (2,400 / 8,000,000)	0.03%	
Incidence ratio		4.33

So indeed, the scarcer factor (GPDH deficiency) has the greater incidence ratio. If we increase the prevalence of GPDH deficiency without changing other parameters, the incidence ratio for fava bean consumption will rise. A spreadsheet is a convenient way to see the effect on incidence ratios from varying the prevalences (check the web page for a downloadable Excel spreadsheet).

Bottom line – what we observe as strength of association is greatly dependent upon prevalence of other component causes.

The above example also illustrates the non-additivity of the attributable risk proportion [ARP=(RR-1)/RR]:

$$\text{ARP for GPDH deficiency} \quad \frac{67.67 - 1}{67.67} = 98.5\ \%$$

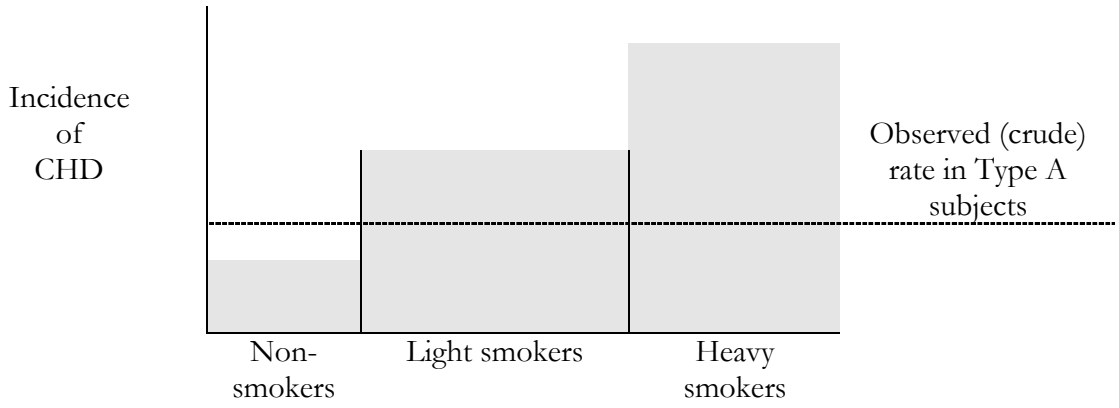
$$\text{ARP for Fava bean consumption} \quad \frac{4.33 - 1}{4.33} = 76.9\ \%$$

Clearly, these ARP's do not sum to 100%, nor, when we think about it, should they.

Before continuing with Rothman's diagrams, we need to revisit an old friend, weighted averages.

Crude rates as weighted averages

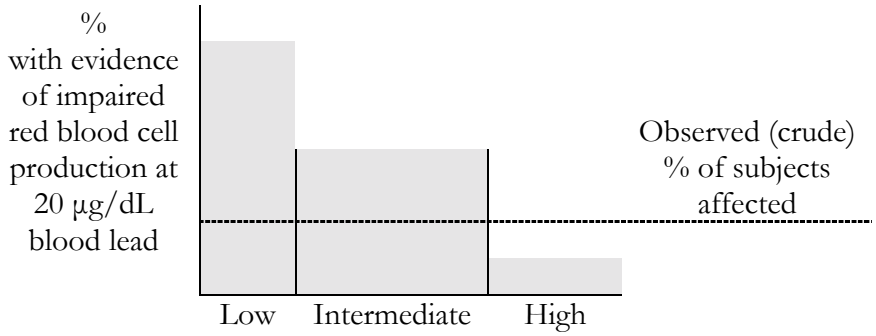
Recall the example of Type A behavior and CHD incidence with which we began the chapter on Confounding. In that example, smokers had a much higher incidence of CHD than did nonsmokers. Since the Type A group consisted mainly of smokers, its CHD incidence was greater than the Type B group, which consisted mainly of nonsmokers. If there were three smoking status groups, then the Type A incidence would be a weighted average of the rates for each of the three smoking status groups (see diagram).



So whenever we compare groups, it is important to pay attention to their distributions of risk factors. In the chapter on confounding, though, we considered only subgroups defined by other (independent) risk factors. We will now see that we must widen our view to include subgroups defined by variables that may influence the effect of the exposure even if those variables have no effect in its absence.

Since every rate we observe in some population is a weighted average of the rates for its component subgroups, this principle must apply to a group of exposed persons as well. Thus, the incidence in the exposed group depends on the composition of the group in regard to factors that are in the same causal pathways as the exposure. A prominent example is genetic factors, which thanks to the molecular biological revolution we are learning a great deal more about.

For example, it has been asserted that susceptibility to impairment of red blood cell production by low-level lead exposure varies according to the genetically-controlled level of the enzyme amino levulanate dehydratase. If that is the case, then in a group of children with a given level of blood lead (e.g., 20 micrograms/dL), the proportion with evidence of impaired red blood cell production would reflect a weighted average of the proportions in each subgroup defined by enzyme level:



Another example is that LDL cholesterol levels reflect both dietary intake of saturated fat and cholesterol and ApoE genotype (from Shpilberg *et al.*, 1997). Compared to persons with the most common allele (E3), those with the E2 allele have lower average cholesterol and those with the E4 allele have higher levels. Therefore, serum cholesterol levels associated with a given population distribution of dietary fat intake will depend on the distribution of these three genotypes.

Yet another example is incidence of venous thromboembolism. A strong effect of oral contraceptives (OC) on venous thromboembolism was one of the first hazards to be recognized for OC. Recent data from Vandenbroucke *et al.* (Factor V Leiden: should we screen oral contraceptive users and pregnant women? *Bio Med J* 1996;313:1127-1130) show an overall incidence of 0.8 per 10,000 women-years that rises to 3.0 per 10,000 women years with OC use. But among OC users who are also carriers of factor V mutation (associated with activated protein C (APC) resistance), the incidence rises to 28.5 per 10,000 women years (from Shpilberg *et al.*, 1997). So the incidence of venous thromboembolism in a population and the effects of OC will be greatly influenced by the population prevalence of factor V mutation.

So whatever phenomenon we are investigating, we need to take account of both independent risk factors for it and factors that may only appear to modify the effect of an exposure of interest (which we will subsequently refer to as an "effect modifier"). This is one reason why we typically stratify data by sociodemographic factors. Factors that affect susceptibility may well covary with demographic characteristics such as age, sex, geographic region, and socioeconomic resources, even if they do not have a role of their own in causation.

Since the distribution of effect modifiers may affect disease rates, it will also affect comparisons between rates in exposed and nonexposed subjects. But if the effect modifier is not itself a risk factor for the disease – i.e., if in the absence of the exposure of interest the effect modifiers is not associated with disease risk – then the modifier can confound associations only among groups with different levels of exposure, not between an exposed and an unexposed group.

Several examples will help to clarify these points. Assume for the moment, that asbestos has no effect on lung cancer incidence independent of smoking, but that smoking has an effect both alone and synergistically with asbestos. A study of the two factors might produce the following data:

Lung cancer rates by smoking and asbestos exposure (per 100,000 person years)

Smokers	
Exposed to asbestos	602
Not exposed to asbestos	123
Nonsmokers	
Exposed to asbestos	11
Not exposed to asbestos	11

From these data we would conclude (leaving aside all issues of statistical significance, bias, and so on) that (a) smoking increases lung cancer risk and (b) asbestos does so only in smokers. Smoking emerges as a risk factor, and asbestos as a modifier of the effect of smoking. Smoking could also be said to be an absolute modifier of the effect of asbestos, since the effect of the latter is null without smoking and dramatic in its presence. The rate ratios for lung cancer in smokers versus nonsmokers are 55 among those exposed to asbestos and 11 among those not exposed.

If we had not analyzed our data separately according to asbestos exposure, the lung cancer rate in nonsmokers would still be 11 per 100,000 person-years. But the rate in smokers would be somewhere between 123 and 602. The actual value would depend on the proportion of smokers exposed to asbestos. Similarly, the rate ratio for lung cancer and smoking would range between 11 and 55. So the crude rate ratio for lung cancer and smoking would always lie within the range of the stratum specific rate ratios.

The fact that the crude rate ratio differs from the stratum-specific rate ratios does not mean that confounding is present. Regardless of the proportion of subjects exposed to asbestos, the relationship between smoking and lung cancer cannot be due to asbestos exposure, though the strength of that relationship will depend on the degree of asbestos exposure. If the crude rate ratio can be expressed as a weighted average of the stratum-specific ratios, then confounding is not present.

The above results will always hold when the effect modifier has no effect in the absence of the exposure and the comparison of interest is between exposed and unexposed groups. A point of theoretical interest is that it was the above type of situation that led us in our discussion of confounding to focus on the question of an association between the potential confounder variable and the disease among the unexposed. An association among the exposed could reflect effect modification rather than independent causation (i.e., among exposed persons, disease rates are higher among those also exposed to a modifier, even if that is not the case among unexposed persons).

Since an effect modifier with no independent effect on the outcome does alter the risk or rate in the presence of exposure, however, an effect modifier can confound comparisons between groups exposed to different degrees. Suppose, for example, that we have divided the smokers in the previous table into light smokers and heavy smokers. Suppose further that most light smokers are exposed to asbestos and most heavy smokers are not. Then we might well observe a higher lung

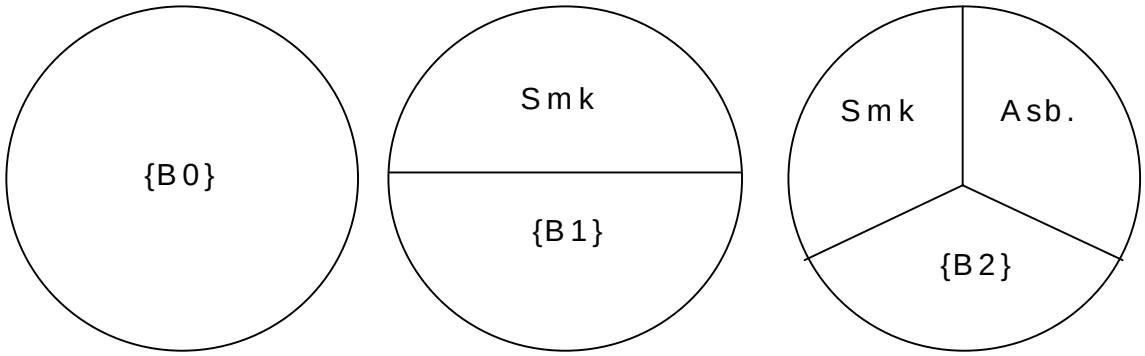
cancer rate among the light smokers (due to their greater asbestos exposure) than among the heavy smokers (where the rate has not been increased by asbestos). The following table gives a numerical illustration of such a situation.

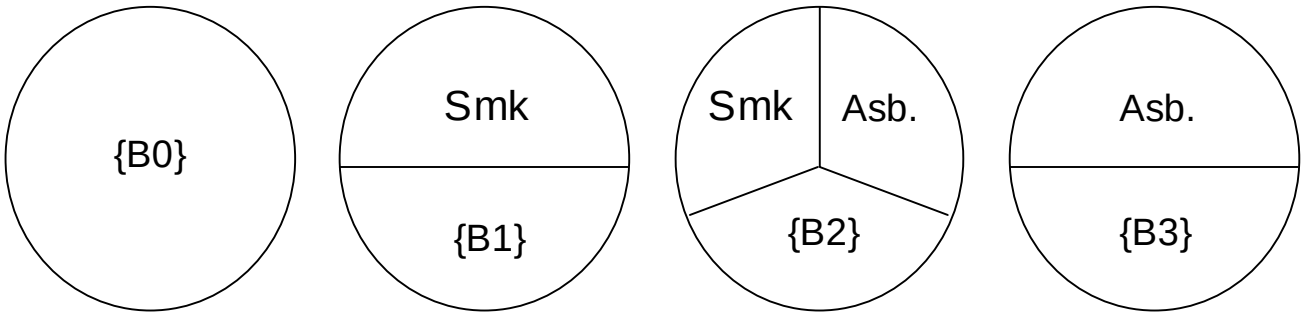
Lung cancer rates by level of smoking and asbestos exposure (per 100,000 person years)

Heavy smokers - overall	(200-1,000)
Exposed to asbestos	1,000
Not exposed to asbestos	200
Light smokers - overall	(100-500)
Exposed to asbestos	500
Not exposed to asbestos	100
Nonsmokers - overall	(11)
Exposed to asbestos	11
Not exposed to asbestos	11

Here, asbestos alone has no effect, heavy smoking in the absence of asbestos has rates twice that for light smoking, and asbestos increases lung cancer rates in smokers fivefold. If 60% of light smokers but only 10% of heavy smokers are exposed to asbestos, then the overall lung cancer rate in light smokers (340 = {500 x .60 + 100 x .40}) will exceed that in heavy smokers (280 = {1,000 x .10 + 200 x .90}).]

While the above situations may at first appear to be complex, they simply reflect different aspects of weighted averages, so with some practice the mystery evaporates. Additional complexity does enter the picture, however, when we turn to effect modification by a variable that has an effect on the outcome by a pathway that does not involve the exposure of interest, i.e., an independent effect. Compare these two causal schema:





where $\{B_0\}$, $\{B_1\}$, $\{B_2\}$, and $\{B_3\}$ are probably overlapping sets of (unidentified) background factors that are needed because (1) people exposed to neither cigarette smoke nor asbestos do get lung cancer, albeit at a low rate; (2) not all people who smoke get lung cancer; etc. [Note: these are the same as Rothman's {U}. I prefer to use different subscripts to make clear that different causal pathways generally involve different required background factors. Otherwise all persons susceptible to developing the disease through a causal pathway involving an exposure (e.g., smoking) would get the disease regardless through the "unexposed" pathway even if not exposed, so the exposure could not be associated with an increased rate of disease.]

The three-pathway configuration represents the situation we have just seen, where asbestos has no effect in nonsmokers. If we apply the data from the preceding numerical example into the upper configuration of causal pathways, we see that the rate that corresponds to the first causal pathway ($\{B_0\}$) is 11/100,000 py. The rate that corresponds to the second causal pathway (Smk|$\{B_1\}$) is 112/100,000 py (123 - 11: the incidence density difference, since people who smoke and can therefore get disease through the second causal pathway are also at risk of developing the disease through the first causal pathway). Similarly, the rate that corresponds to the third causal pathway (Smk|Asb|$\{B_2\}$) is (602-112-11)/100,000 py (since we observe 602 for people who have both exposures, but they could have developed disease from either of the first two causal pathways). These different disease rates presumably correspond to the prevalences of $\{B_1\}$, $B_2\}$, and $\{B_3\}$.

The four-pathway configuration does show an <u>independent effect</u> of asbestos. In this configuration, we see that confounding by asbestos can occur, since the risk in nonsmokers may be elevated by the effect of asbestos. Moreover, it now becomes more difficult to assess effect modification as "a combined effect greater than we expect from the effects of each variable acting alone". The problem is: if each variable has an effect on its own, what do we expect for their combined effect so we can say whether we have observed something different from that?

Consider, for example, actual data on the relationship of smoking and asbestos to lung cancer death rates (from E. Cuyler Hammond, Irving J. Selikoff, and Herbert Seidman. Asbestos exposure, cigarette smoking and death rates. *Annals NY Acad Sci* 1979; 330:473-90).

Age standardized lung cancer rates by smoking and asbestos exposure (per 100,000 person years)

	Smokers	Nonsmokers
Exposed to asbestos	602	58
Not exposed to asbestos	123	11

When we calculate the disease rates that correspond to each of the four causal pathways in the lower configuration of causal "pies" above, the two leftmost pathways have the same rates as in the upper configuration. The rate corresponding to the rightmost pathway (Asbestos|{B3}) is 58-11 = 47/100,000 py. The rate that corresponds to the third causal pathway (Smk|Asb|{B2}) is now reduced since some of cases with both exposures could be due to the effect of asbestos. So the rate that corresponds to the third pathway is now (602-112-11-47)/100,000 py = 410/100,000 py.

We might take these rates and reason as follows:

Increase due to smoking	123 - 11 = 112
Increase due to asbestos	58 - 11 = 47
Total increase expected due to both	112 + 47 = 159
Total observed increase	602 - 11 = 591 !

Since the increase due to the combined effect greatly exceeds that expected from our (additive) model, we would conclude that the effect is synergistic.

Alternatively, we might reason in relative terms:

Relative increase due to smoking	123 / 11 = 11.2
Relative increase due to asbestos	58 / 11 = 5.3
Total increase expected due to both	11.2 x 5.3 = 59.4
Total observed increase	602 / 11 = 54.7

This time the observed increase and that expected from our (multiplicative) model are quite close, so we conclude that there is no effect modification. We are thus faced with a situation where the decision about effect modification depends upon what model we employ to arrive at an expected joint effect to compare with the observed joint effect (or equivalently, upon the scale of measurement, hence the term "effect measure modification").

Before pondering this dilemma further, we should first state the additive and multiplicative models explicitly. To do so we introduce a notation in which "1" indicates presence of a factor, a "0" indicates absence of a factor, the first subscript represents the first risk factor, and the second subscript represents the second risk factor (see below).

Notation for joint effects

R_1	risk or rate in the presence of a factor, ignoring the presence or absence of other identified factors
R_0	risk or rate in the absence of a factor, ignoring the presence or absence of other identified factors
R_{11}	risk or rate when both of two factors are present
R_{10}	risk or rate when the first factor is present but not the second
R_{01}	risk or rate when only the second factor is present
R_{00}	risk or rate when neither of the two factors is present (i.e., risk due to background factors)
RD_{11}	difference between the risk or rate when both factors are present and the risk or rate when neither factor is present
RD_{10}	difference between the risk or rate when only the first factor is present and the risk or rate when neither factor is present
RD_{01}	difference between the risk or rate when only the second factor is present and the risk or rate when neither factor is present
RR_{11}	ratio of the risk or rate when both factors are present divided by the risk or rate when neither factor is present
RR_{10}	ratio of the risk or rate when only the first factor is present divided by the risk or rate when neither factor is present
RR_{01}	ratio of the risk or rate when only the second factor is present divided by the risk or ratio when neither factor is present

The use of two subscripts implies a stratified analysis. The first subscript indicates presence or absence of the first factor; the second subscript, presence or absence of the second factor. For example, R_{10} refers to the rate for persons exposed to the first factor but not to the second. That rate can be referred to as the rate for the exposed (to factor 1) in the stratum without factor 2; equivalently, R_{10} can be referred to as the rate for the unexposed (to factor 2) in the stratum where factor 1 is present. In contrast, a single subscript (R_1) means the factor is present, with other factors present or not present (i.e., crude with respect to other factors). "Background" factors and the risk R_{00} associated with them are assumed to be uniformly distributed across all strata.

Additive model

Under an additive model, the increase in rate or risk from a combination of factors equals the sum of the increases from each factor by itself. We can express this statement algebraically, using the rate (or risk) difference:

$$R_{11} - R_{00} = R_{10} - R_{00} + R_{01} - R_{00} \quad \text{(A1)}$$

$$RD_{11} = RD_{10} + RD_{01} \quad \text{(A2)}$$

Using elementary algebra and the definition of the rate difference, we can also write the additive model as:

$$R_{11} = R_{00} + RD_{10} + RD_{01} \quad \text{(A3)}$$

i.e., the expected rate where both factors are present is the baseline rate (R_{00}, neither factor present) plus the rate difference associated with the first factor plus the rate difference associated with the second factor. Another equivalent expression is:

$$R_{11} = R_{10} + R_{01} - R_{00} \quad \text{(A4)}$$

Since $RR_{11} = R_{11}/R_{00}$, $RR_{10} = R_{10}/R_{00}$, and $RR_{01}=R_{01}/R_{00}$, we can express the additive model in terms of the risk (or rate) ratio, by dividing each term in expression A1 by the baseline risk, R_{00}.

$$RR_{11} - 1 = RR_{10} - 1 + RR_{01} - 1 \quad \text{(A5)}$$

An advantage of this formulation is that we can use it even when we do not have estimates of specific risks or risk differences. The expression $(R_1-R_0)/R_0$, or RR – 1, is sometimes referred to as the (relative) **excess risk**. The additive model, expressed in terms of excess risk, is therefore:

Excess risk for A and B together = Excess risk for A + Excess risk for B

i.e., the joint excess risk equals the sum of the excess risk for each factor alone. With this expression we can evaluate the additive model even from case-control data.

More than two factors

Where there are three factors, we have, analogously:

$$RR_{111} - 1 = RR_{100} - 1 + RR_{010} - 1 + RR_{001} - 1 \quad \text{(A6)}$$

$$RD_{111} = RD_{100} + RD_{010} + RD_{001} \quad \text{(A7)}$$

$$R_{111} = R_{000} + RD_{100} + RD_{010} + RD_{010} \quad \text{(A8)}$$

and

$$R_{111} = R_{100} + R_{010} + R_{001} - 2\,R_{000} \quad \text{(A9)}$$

So the additive model can be regarded as based on 1) additivity of excess risks, 2) additivity of risk differences, and/or 3) additivity of the risks themselves. The reason that we need to subtract the baseline risk in the last of these forms is that risk in the presence of any of the factors includes, necessarily, the ever-present background risk. So when we add the risk for one factor to the risk for another factor, the background risk is added twice. Thus, when we refer to R_{ijk} as the risk (or rate) for a factor "by itself", the **"by itself" really means "with no other specified factors"**, since the **baseline risk is, by definition, always present**.

Multiplicative model

In parallel fashion, the multiplicative model assumes that the relative risk (risk ratio, rate ratio) for the factors operating together equals the product of their relative risks:

$$RR_{11} = RR_{10} \times RR_{01} \quad \text{(M1)}$$

Multiplying through by baseline risk (R_{00}) gives:

$$R_{11} = R_{00} \times RR_{10} \times RR_{01} \quad \text{(M2)}$$

and

$$R_{11} = R_{10} \times R_{01} / R_{00} \quad \text{(M3)}$$

i.e., the joint risk equals the product of 1) the baseline risk multiplied by the relative risk for each factor and/or 2) the individual risks and the reciprocal of the baseline risk. For three factors, the model becomes:

$$RR_{111} = RR_{100} \times RR_{010} \times RR_{001} \quad \text{(M4)}$$

and

$$R_{111} = R_{000} \times RR_{100} \times RR_{010} \times RR_{001} \quad \text{(M5)}$$

and

$$R_{111} = R_{100} \times R_{010} \times R_{001} / (R_{000})^2 \qquad \text{(M6)}$$

Again, there is a baseline risk or rate in the denominator of each relative risk, so when the relative risks are converted to risks, the R_{000} in the numerator eliminates one of the resulting three R_{000}'s, leaving two remaining in the denominator. As before, "by itself" means without other specified factors, but including baseline risk.

Note, however, that the multiplicative model can also be written as an additive model on the logarithmic scale (because addition of logarithms is equivalent to multiplication of their arguments):

$$\ln(R_{111}) = \ln(R_{100}) + \ln(R_{010}) + \ln(R_{001}) - 2 \times \ln(R_{000}) \qquad \text{(M7)}$$

For this reason, the difference between the additive and multiplicative models can be characterized as a transformation of scale. So "effect modification" is scale-dependent.

Optional aside – It can also be shown that a multiplicative model can be expressed as an additive model on the natural scale plus an interaction term. For two factors: $(R_{10} - R_{00})(R_{01} - R_{00})/R_{00}$, or equivalently, $(R_{00})(RR_{10}-1)(RR_{01}-1)$ – essentially, we add a "fudge factor".

Additive model:

$$R_{11} = R_{10} + R_{01} - R_{00}$$

Additive model with interaction term:

$$R_{11} = R_{10} + R_{01} - R_{00} + R_{00} \times (RR_{10}-1) \times (RR_{01}-1)$$

Multiplying out the interaction term:

$$R_{11} = R_{10} + R_{01} - R_{00} + R_{00} \times RR_{10} \times RR_{01} - R_{00} \times RR_{10} - R_{00} \times RR_{01} + R_{00}$$

Dividing both sides by R_{00}:

$$RR_{11} = RR_{10} + RR_{01} - 1 + RR_{10} \times RR_{01} - RR_{01} - RR_{10} + 1$$

Simplifying:

$$RR_{11} = RR_{10} \times RR_{01} = \text{the multiplicative model}$$

[End of aside]

The choice of model – additive, multiplicative, or other – is not a settled affair and involves a variety of considerations. One consideration is to choose the simplest model that can represent the data. Recall the example from an earlier lecture:

Relative versus Absolute Effects example Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years

Age	Cigarettes/day	OC*	$\overline{OC}$*	RR**	AR***
30-39	0-14	6	2	3	4
	15 +	30	11	3	19
40-44	0-14	47	12	4	35
	15 +	246	61	4	185

Notes:
* Rate per 100,000 women-years
** RR=relative risk (rate ratio)
*** AR=attributable risk (rate difference)
Source: Mann *et al.* (presented in a seminar by Bruce Stadel)

Here, we saw that the rate ratio was a more stable index of the strength of association between OC and MI across the various combinations of age and smoking. In fact, the MI rates for many combinations of the three risk factors – age, smoking, and OC – are not far from those expected based on the multiplicative model. To see this, use the additive and multiplicative models just presented with the data in the above table to fill in the rightmost two columns of the following table. If we write the rates for the three risk factors as R_{100}, R_{010}, and R_{001}, with the background rate defined as R_{000}, then joint rates for several combinations of risk factors would be:

First and third factors present (row 6):

$$R_{101} = R_{100} + R_{001} - R_{000} \qquad \text{(additive model)}$$

$$R_{101} = R_{100} \times R_{001} / R_{000} \qquad \text{(multiplicative model)}$$

First and second factors present (row 7):

$$R_{110} = R_{100} + R_{010} - R_{000} \qquad \text{(additive model)}$$

$$R_{110} = R_{100} \times R_{010} / R_{000} \qquad \text{(multiplicative model)}$$

All three factors present (row 8):

$$R_{111} = R_{100} + R_{010} + R_{001} - 2\,R_{000} \quad \text{(additive model)}$$

$$R_{111} = R_{100} \times R_{010} \times R_{001} / (R_{000})^2 \quad \text{(multiplicative model)}$$

where the three factors are 1) age, 2) cigarette smoking, and 3) oral contraceptives. For example, suppose R_{101} is the rate of MI in women who are in the older age group, smoke less than 15 cigarettes/day or not at all, and use oral contraceptives. The multiplicative model says that the rate for any combination of the three factors (with cutpoints defined as in the table) equals the product of the rates for each of the three factors when neither of the other two is present, divided by the square of the rate for those who have none of the three factors (i.e., only unidentified background factors are present). Here is a "test" of the model (one line is left incomplete, to give you the satisfaction of figuring it out):

Home-made multiplicative model of Incidence of myocardial infarction (MI) in oral contraceptive (OC) users per 100,000 women-years

Row		Age	Cigarettes /day	OC*	Observed Rate	Expected (Multiplic)	Expected (Additive)
1	R_{000}	0: 30-39	0: 0-14	0: no	2	-	-
2	R_{001}	0: 30-39	0: 0-14	1: yes	**6**	-	-
3	R_{010}	0: 30-39	1: 15 +	0: no	**11**	-	-
4	R_{100}	1: 40-44	0: 0-14	0: no	**12**	-	-
5	R_{011}	0: 30-39	1: 15 +	1: yes	30	___	___
6	R_{101}	1: 40-44	0: 0-14	1: yes	47	36	16
7	R_{110}	1: 40-44	1: 15 +	0: no	61	66	21
8	R_{111}	1: 40-44	1: 15 +	1: yes	246	198	25

Notes: 0: and 1: indicate the coding for each risk factor level. Rates for single factors in the absence of the other two are shown in bold.

[Thanks to Jim McDougal (1996) for spotting my longstanding errors in the 3-factor interaction in this table and its explanation.]

Certainly the multiplicative model yields expected rates that are closer to the observed rates for various combinations of the factors than does the additive model. The better fit for the

multiplicative model supports the use of the rate ratio as the measure of association for each risk factor and each risk factor combination in these data. If Mann *et al*. want a summary measure for the effect of OC on MI rates, controlling for age and smoking, a weighted average of the rate ratios (3, 3, 4, 4) for OC use across the four age and smoking categories would be a good choice. But then what happened to effect modification?

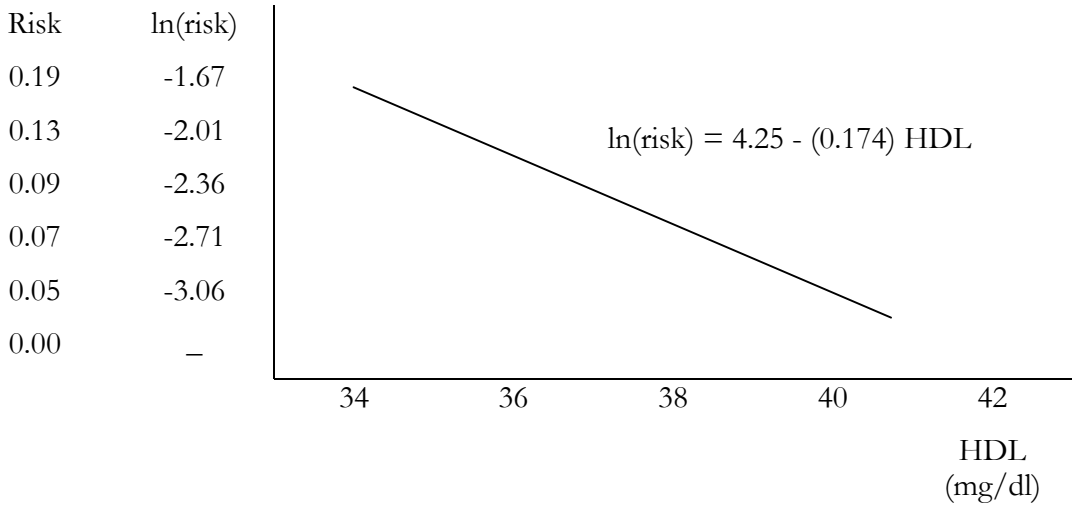
The "natural" scaling

The additive model has been put forth by Rothman as the "natural" scaling. Risks are probabilities, and the probability that either of two independent and mutually exclusive events will take place (e.g., smoking causes MI or OC causes MI) is the sum of the probabilities for each. Therefore if the risk (probability of disease) in people with both exposures exceeds the sum of the risks for each exposure separately, then some non-independence (i.e., interaction) must exist between these two disease events. Rothman's proposition appears to have become the consensus in terms of evaluating impact on public health and/or individual risk (see below). Our earlier suggestion that the risk or rate difference serves more often as a measure of impact than as a measure of strength of association in respect to etiology is distinctly parallel.

When our interest is the relationship of the mathematical model or scaling to possible biological mechanisms, however, the issue becomes more problematic. Kupper and Hogan (Interaction in epidemiologic studies. *Am J Epidemiol* 108:447-453, 1978) demonstrated how two factors having biologically equivalent modes of action, so that either factor can be regarded as a different concentration of the other, can appear to be synergistic in their joint effect if the dose-response curve is nonlinear. (This example harks back to the fact that additivity on the logarithmic scale is equivalent to multiplicativity on the natural scale.) Therefore, a departure from additivity can occur even in the absence of biological interaction.

Data from a study published in that year provides an illustration. Bradley DD, *et al*. (Serum high-density-lipoprotein cholesterol in women using oral contraceptives, estrogens and progestins. *New Engl J Med* 299:17-20, 1978) suggested that smoking and oral contraceptives (OC) may each increase myocardial infarction risk by reducing levels of HDL cholesterol. The effects of smoking and oral contraceptives on HDL appear to be additive. But if the relationship between HDL level and myocardial infarction risk is exponential, with the logarithm of risk increasing in linear fashion with declining HDL, then the effects of the two behavioral risk factors on myocardial infarction risk will be multiplicative.

In the figure below, the natural logarithm of heart attack risk is a linear function of HDL level, so that risk rises exponentially as HDL decreases. The risk function comes from Bradley *et al*.'s paper.



If smoking causes a reduction in HDL of 6 mg/dL, and oral contraceptives cause a reduction of 2 mg/dL, then the changes in ln(risk) [from the formula in the figure] and in the RR's for smoking and oral contraceptives separately and for both together are shown in the following table:

Factors	**HDL reduction**	**Increase in ln(risk)**	**RR**
Smoking	6	1.044	2.84
OC only	2	0.348	1.42
Smoking and OC	8	1.392	4.02

Smoking is associated with a 6 mg/dL lower HDL level, corresponding to an increase in ln(risk) of 1.044, which in turn corresponds to a relative risk of 2.84. Although (in this conceptual model) the biological effects of smoking and OC on HDL are additive, because the dose-response curve is not linear, this additivity of dose does not imply additivity of response.

This point has been elaborated by Thompson (1991), who makes the point that pathogenetic processes are likely to include factors that intervene between the variables in our simplified causal models. Such intervening factors are generally unknown or unmeasured by epidemiologists. Yet as illustrated above, the form of the functional relation between two variables can change the appearance of a risk function. The actions of two factors may be additive on their immediate target, but their effect on risk of a downstream effect could be additive, multiplicative, or anything else. Only in the case of a crossover effect (a.k.a. qualitative interaction, which to be certain that it exists should be demonstrated by confidence intervals that lie wholly below the null value in one stratum and wholly above the null value in the other stratum – see Thompson 1991) do we have a basis for inferring that something of biological interest is occurring (after excluding other non-mathematical explanations). Another situation where interpretation is unambiguous – what I have called "absolute effect modification", where one factor has no effect in the absence of the other – is in practice just as problematic as other non-crossover situations, since it is rarely possible to exclude the presence of at least a weak effect (Thompson 1991).

Effect modification as a reflection of information bias:

Another consideration that arises in interpreting apparent effect modification in epidemiologic data relates to the question of the actual dosage received by subjects. Suppose that data from a study of lung cancer and smoking yielded these results:

Lung cancer rates per 100,000 person-years

	Males	**Females**
Smokers	300	500
Nonsmokers	50	50

The rate ratios for males and females are 6 (300/50) and 10 (500/50), respectively, which might suggest that women are more susceptible to the carcinogenic properties of tobacco smoke. But what if women smokers inhale more deeply and therefore receive a larger dose of carcinogenic substances, the actual exposure? So whereas effect measure modification in epidemiologic data may suggest the need for a more detailed understanding of the phenomenon under study, an interpretation in terms of biological synergism involves causal inference and needs to be approached from that perspective.

Consensus

Rothman, Greenland, and Walker (1980) presented four perspectives on the concept of interaction:

1. The biologic perspective is concerned with elucidating how various factors act at the biological (mechanistic) level.
2. The statistical perspective treats interaction as "leftovers", i.e., the nonrandom variability in data that is not accounted for by the model under consideration. Statisticians often try to reformulate the model to eliminate these leftovers, i.e., to find the simplest model that fits the data adequately.
3. The public health perspective should regard interaction is a departure from additivity, if one assumes that costs are proportional to the number of cases. If effects are more than additive, then a greater than proportional payoff can be obtained by intervening against a factor involved in an interaction.
4. The individual decision-making perspective should also regard interaction as a departure from additivity, again assuming a linear relationship between costs and, in this case, risk. For example, if the combined effect of smoking and hypertension on CHD risk is greater than additive, someone with hypertension can reduce his risk even more by quitting smoking than someone with normal blood pressure.

These perspectives appear to be widely accepted. The term "effect modification" is generally used to refer to a meaningful departure from a given mathematical model (i.e., additive, multiplicative, or whatever) of how risks or rates combine. ("Meaningful" means that the departure is large enough to

have clinical or public health significance and thought not to be due to random variability, measurement inadequacy, or confounding.) The additive model appears to be accepted as the indicator of "expected joint effects" for policy or decision-making considerations.

Summary

In view of the foregoing, we may attempt to summarize the relevance of interaction and effect modification in terms of four implications:

1. Increasing the precision of description and prediction of phenomena under study;
2. Indicating the need to control for the factors that appear as modifiers;
3. Suggesting areas for developing etiologic hypotheses; and
4. Defining subgroups and factor combinations for special attention for preventive approaches.

Elaboration

1. Increasing precision of description:

In our smoking in men and women illustration, the different strength of the smoking-lung cancer association between men and women may lead to an appreciation of the need to be more precise in the measurement and specification of the exposure variable.

2. Indicating the need to control for modifiers:

Since an effect modifier changes the strength of the association under study, different study populations may yield different results concerning the association of interest. Unlike potential confounders, modifying variables cannot create the appearance of an association (for exposed versus unexposed) where none exists. But the proportion of the study population that has a greater susceptibility will influence the strength of the association. Therefore, to achieve comparability across studies, it is necessary to control for the effect of the modifying variables, generally by carrying out a separate analysis at each level of the modifier.

3. Developing etiologic hypotheses:

Attention to interactions in the data may lead to the formulation of etiologic hypotheses that advance our understanding of the pathogenetic processes involved. Although the linkage between mechanisms and relationships in data is uncertain, a strong interaction might suggest that a shared mechanism is involved. For example, the interaction of smoking and asbestos might suggest a scenario such as impairment of lung clearing processes and/or of mechanical injury from asbestos particles increases susceptibility to carcinogens in cigarette smoke.

4. Defining subgroups for preventive approaches:

To observe that the OC-MI association is particularly strong among smokers and/or women over 35 carries evident preventive implications in terms of health education warnings, contraindications to prescribing, targeting of messages, and so forth. The synergistic relationship between smoking and asbestos in the etiology of lung cancer suggests the value of extra efforts to convince asbestos workers not to smoke. If the cost of helping a smoker to quit smoking is the same for asbestos workers and others, then the benefit-cost ratio will be greater for a cessation program with smokers who work with asbestos because more cases of lung cancer will be avoided for the same number of quitters.

The rationale for viewing effect modification as a departure from an additive model of disease risks, at least for public health purposes, is that if an additive model holds, then removal of one agent can only be expected to eliminate the risk that arises from that agent but not the risk from other agents. If there is positive interaction, however, removal of any one of the agents involved will reduce some risk resulting from the other as well. In such a situation, the impact of removing a risk factor is greater than that expected on the basis of its effect on baseline risk.

A "real-life" example

The following table comes from a randomized, controlled trial of a self-help smoking cessation intervention using brief telephone counseling. Quit rates for smokers in the intervention group and the other groups suggested that participants with certain baseline characteristics were more or less likely to benefit from the telephone counseling intervention. For example, the telephone counseling intervention was associated with a 14 percentage point (31%–17%) higher quit rate for participants who were not nicotine dependent but with only a 3 percentage point (17%–14%) higher quit rate for participants who were nicotine dependent. The intervention was associated with a 12 percentage point (29%-17%) higher quit rate for participants who had not previously undergone an intensive cessation program but with only a 2 percentage point (17%–15%) higher quit rate for participants who had. The observed differences appeared to be consistent with the fact that the intervention was a minimal treatment (so would not be of much help to a smoker who had already experienced an intensive treatment program) that incorporated nicotine-fading/brand-switching (which has limited applicability for a smoker who is already smoking a low-nicotine brand).

Baseline Characteristics Associated with a Significantly Different Telephone Counseling Effect on 7-day Abstinence at 16-months Follow-up in 1,877 Smokers at Group Health Cooperative of Puget Sound, Washington, 1985-1987

	Quit rate			
	with characteristic		without characteristic	
Baseline characteristic	**Counseling**	**No Counseling**	**Counseling**	**No Counseling**
Nicotine dependent	17	14	**31**	**17**
Intensive treatment	17	15	**29**	**17**
Brand nicotine > 0.7 mg	**24**	**12**	22	20
VIP better role model	**28**	**15**	19	16
Close friends/relatives	21	17	**29**	**14**
Nonsmoking partner	19	19	**25**	**14**

Note: For each characteristic, the difference in quit rates between counseling and no-counseling groups among those with the characteristic is significantly ($p<0.05$) greater or less (by about 10 percentage points) than the quit rate difference among those without the characteristic. Bolding denotes the greater telephone counseling effect.

Reference: Schoenbach VJ, Orleans CG, Wagner EH, Quade D, Salmon MAP, Porter CQ. Characteristics of smokers who enroll and quit in self-help programs. *Health Education Research* 1992;7:369-380, Table 3.

Bibliography

Rothman and Greenland, *Modern epidemiology*.

Hertz-Picciotto I, Neutra RR. Resolving discrepancies among studies: the influence of dose on effect size. *Epidemiology* 1994;5:156-163.

Koopman, James S. and Douglas L. Weed. Epigenesis theory: a mathematical model relating causal concepts of pathogenesis in individuals to disease patterns in populations. *Am J Epidemiol* 1990; 132:366-90.

Marshall, Roger J. Confounder prevalence and stratum-specific relative risks: implications for misclassified and missing confounders. *Epidemiology* 1994;5:439-44

Koopman, J.S.: Interaction between discrete causes. *Am J Epidemiol* 113:716-724, 1981.

Khoury, Muin J.; W. Dana Flanders, Sander Greenland, Myron J. Adams. On the measurement of susceptibility in epidemiologic studies. *Am J Epidemiol* 1989; 129:183-90.

Rothman, KJ. Causes. *Am J Epidemiol* 1976; 104:587-92.

Rothman, K.J.: Synergy and antagonism in cause-effect relationships. *Am J Epidemiol* 99:385-388, 1974.

Rothman, K.J., Greenland, S., and Alexander M. Walker: Concepts of interaction. *Am J Epidemiol* 112:467-470, 1980.

Shpilberg O, Dorman JS, Ferrell RE, Trucco M, *et al*. The next stage: Molecular epidemiology. *J Clin Epidemiol* 1997;50:633-638.

Siemiatycki, Jack and Duncan C. Thomas. Biological models and statistical interactions: an example from multistage carcinogenesis. *International J Epidemiol* 10:383-387, 1981.

Thompson, W. Douglas. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;44:221-232.

Walker, Alexander M. Proportion of disease attributable to the combined effect of two factors. *International J Epidemiology* 1981; 10:81-85.

Weed, Douglas L.; Michael Selmon, Thomas Sinks. Links between categories of interaction. *Am J Epidemiol* 1988; 127:17-27.

Weiss, Noel S. Accounting for the multicausal nature of disease in the design and analysis of epidemiologic studies. *Am J Epidemiol* 1983;117:14-18.