11. Multicausality: Confounding

Accounting for the multicausal nature of disease – secondary associations and their control

Introduction

When "modern epidemiology" developed in the 1970s, Olli Miettinen organized sources of bias into three major categories: selection bias, information bias, and confounding bias. If our focus is the crude association between two factors, selection bias can lead us to observe an association that differs from that which exists in the population we believe we are studying (the target population). Similarly, information bias can cause the observed association to differ from what it actually is. Confounding differs from these other types of bias, however, because confounding does not alter the crude association. Instead, concern for confounding comes into play for the interpretation of the observed association.

We have already considered confounding, without referring to it by that term, in the chapter on age standardization. The comparison of crude mortality rates can be misleading, not because the rates are biased, but because they are greatly affected by the age distributions in the groups being compared. Thus, in order to be able to interpret the comparison of mortality rates we needed to examine age-specific and age-standardized rates in order avoid or equalize the influence of age. Had we attemped to interpret the crude rates, our interpretation would have been ***confounded*** by age differences in the populations being compared. We therefore ***controlled for*** the effects of age in order to remove the confounding. In this chapter we will delve into the mechanics of confounding and review the repertoire of strategies to avoid or control it.

Counterfactual reasoning

Epidemiologic research, whether descriptive or analytic, etiologic or evaluative, generally seeks to make causal interpretations. An association between two factors prompts the question what is responsible for it (or in the opposite case, what is responsible for our not seeing an association we expect). Causal reasoning about associations, even those not the focus of investigation, is part of the process of making sense out of data. So the ability to infer causal relationships from observed associations is a fundamental one.

In an "epidemiologists' ideal world", we could infer causality by comparing a health outcome for a person exposed to a factor of interest to what the outcome would have been in the absence of exposure. A comparison of what would occur with exposure to what would occur in the absence of exposure is called counterfactual, because one side of the comparison is contrary to fact (see Rothman and Greenland, p49, who attribute this concept to Hume's work in the 18th century). This counterfactual comparison provides a sound logical basis for inferring causality, because the effect of the exposure can be isolated from the influence of other factors.

In the factual world, however, we can never observe the identical situation twice, except perhaps for "instant replay", which does not allow us to alter exposure status. The plethora of factors that can influence an outcome vary from person to person, place to place, and time to time. Variation in these factors is responsible for the variability in the outcomes we observe, and so a key objective in both experimental and observational research is to minimize all sources of variability other than the one whose effects are being observed. Only when all other sources of variability are adequately controlled can differences between outcomes with and without the exposure be definitively attributed to the exposure.

Experimental sciences

Experimental sciences minimize unwanted variability by controlling relevant factors through experimental design. The opportunities for control that come from laboratory experimentation are one of the reasons for their power and success in obtaining repeatable findings. For example, laboratory experiments can use tissue cultures or laboratory animals of the same genetic strain and maintain identical temperature, lighting, handling, accommodation, food, and so forth. Since not all sources of variability can be controlled, experiments also employ control groups or conditions that reflect the influence of factors that the experimenter cannot control. Comparison of the experimental and control conditions enables the experimenter to control analytically the effects of these unwanted influences.

Because they can manipulate the object of study, experiments can achieve a high level of assurance of the equivalence of the experimental and control conditions in regard to all influences other than the exposure of interest. The experimenter can make a before-after comparison by measuring the outcome before and after applying an "exposure". Where it is important to control for changes that occur with time (aging), a concurrent control group can be employed. With randomized assignment of the exposure, the probability of any difference between experimental and control groups can be estimated and made as small as desired by randomizing a large number of participants. If the exposure does not have lingering effects, a cross-over design can be used in which the exposure is applied to a random half of the participants and later to the other half. The before-after comparison controls for differences between groups, and the comparison across groups controls for changes that occur over time. If measurements can be carried out without knowledge of exposure status, then observer effects can be reduced as well. With sufficient control, a close approximation to the ideal, counterfactual comparison can be achieved.

Comparison groups

In epidemiology, before-after and cross-over studies are uncommon, partly because the exposure often cannot be manipulated by the investigator; partly because of the long time scale of the processes under study; and partly because either the exposure, the process of observation, or both often have lasting effects. The more usual approximation to a counterfactual comparison uses a comparison group, often called a "control group" on analogy with the experimental model. The comparison group serves as a surrogate for the counterfactual "exposed group without the exposure". Thus, the adequacy of a comparison group depends upon its ability to yield an accurate estimate of what the outcomes would have been in the exposed group in the absence of the exposure.

Randomized trials

The epidemiologic study design that comes closest to the experimental model is the large randomized, controlled trial. However, the degree of control attainable with humans is considerably less than with cell cultures. For example, consider the Physicians Health Study, in which Dr. Charles Hennekins and colleagues at Harvard University enrolled U.S. physicians (including several faculty in my Department) into a trial to test whether aspirin and/or beta carotene reduce risk of acute myocardial infarction and/or cancer. The study employed a factorial design in which the physicians were asked to take different pills on alternate days. One group of physicians alternated between aspirin and beta carotene; another group alternated between aspirin and a placebo designed to look like a beta carotene capsule; the third group alternated between an aspirin look-alike and beta carotene; and the fourth group alternated between the two placebos. In this way the researchers could examine the effects of each substance both by itself and with the other – two separate experiments conducted simultaneously.

With 20,000 participants, this study design ensured that the four groups were virtually identical in terms of baseline characteristics. But there was clearly less control over physicians during the follow-up period than would have been possible with, say, laboratory rats. For example, the physician-participants may have increased their exercise levels, changed their diets, taken up meditation, or made other changes that might affect their disease risk. Such changes can render a study uninformative.

The MRFIT debacle

Just such an unfortunate situation apparently developed in the Multiple Risk Factor Intervention Trial (MRFIT), a large-scale (12,000 participants, over $100 million) study sponsored by the National Heart, Lung, and Blood Institute (NHLBI) of the U.S. National Institutes of Health (NIH). As evidence mounted that blood cholesterol was an etiologic risk factor for multiple forms of cardiovascular disease, particularly coronary heart disease (CHD), the possibility for a trial to verify that changing cholesterol levels would reduce CVD was being intensively explored. However, in the late 1960's suitable drugs were not available; the only cholesterol-lowering intervention was dietary modification. A "diet-heart" trial would require over one million participants and last many years – not an appealing scenario.

The idea of a diet-heart trial persisted, however, eventually metamorphosizing into a study to verify that cardiovascular disease rates could be lowered by changing the three most common CVD risk factors: cigarette smoking, elevated serum cholesterol, and hypertension. Thus was born MRFIT.

The trial was launched in the early 1970's. Men (because they have higher CHD rates) whose risk factors placed them at high CHD risk (based on a model from the Framingham Study) were randomized to "Special Intervention" (SI) or Usual Care (UC). SI participants received intensive, state-of-the-art, theoretically-based interventions to improve diet and promote smoking cessation.

Hypertensive SI participants were treated with a systematic protocol to control their blood pressure. UC participants had copies of their regular examinations sent to their personal physicians, but received no treatment through MRFIT. In this pre-"wellness" (health promotion / disease prevention through individual behavior change) era, the trial's designers projected modest risk factor changes in SI participants and little if any change in UC participants. Even though UC participants' physicians were to receive examination results, in those years few practicing physicians became involved in dietary change, smoking cessation, or even blood pressure control for healthy patients.

The planned sample size of about 12,000 men, about 6,000 in SI and 6,000 in UC, was achieved, and follow-up was maintained for seven years. By the end of the follow-up period, risk factor levels in the SI group had reached the target levels, and 46% of SI smokers quit smoking. But to the surprise (and consternation) of the MRFIT investigators, cholesterol levels and blood pressures also declined among UC participants, and 29% of UC smokers quit. During the years of the trial, smoking, diet, and hypertension had risen on the agendas of both the medical profession and the public (presumably aided by another NHLBI initiative, the National High Blood Pressure Control Program). Mortality among the UC participants was not only considerably lower than the projection based on data from the Framingham study but was even (slightly) below that for SI participants. Needless to say, there were many uncomfortable epidemiologists when the results came out.

Nonrandomized studies

Most epidemiologic studies do not have the opportunity to compare groups formed by a random assignment procedure. Whether we study smoking, alcohol, seat belts, handgun ownership, eating, exercise, overweight, use of particular medications, exposure to toxic agents, serum cholesterol, blood pressure, air pollution, or whatever, there is no assurance that the comparison group (the unexposed participants) is just like the exposed participants except for the exposure under study. Indeed, the opposite is more likely, since all sorts of factors are related to family and physical environment, occupation (e.g., workplace exposures), lifestyles (e.g., nutrition, physical activity), social influences (e.g., social support, injustice), health care, health conditions (e.g., medications), genetic endowment, and other characteristics.

Confounding

Thus, whenever we compare groups with respect to factors of interest, we must always consider that group differences in other, "extraneous" factors could be responsible for what we observe (or do not observe) (extraneous factors = factors other than the relationships under study). ***Confounding*** (from the Latin *confundere*, to mix together) can be defined as a "situation in which a measure of the effect of an exposure on risk is distorted because of the association of exposure with other factor(s) that influence the outcome under study" (Last, *A dictionary of epidemiology*). Confounding is a problem of comparison, a problem that arises when extraneous but important factors are differently distributed across groups being compared. The centrality of the concept of confounding and its control in epidemiology derives from the limited opportunities for experimental control.

A hypothetical example (with apologies to the Western Collaborative Group Study)

To investigate how confounding can arise and how it can be dealt with, consider the following hypothetical data based on the Western Collaborative Group Study of coronary heart disease (CHD) risk in managers and white collar workers exhibiting the coronary prone behavior pattern. This pattern, most often referred to as the Type A behavior pattern, is described as hard-driving, time-urgent, and hyperaggressive. In contrast, Type B people are regarded as more relaxed and easy-going.

In this study, Meyer Friedman, Raymond Rosenman, and their colleagues recruited 3,154 white male managers, aged 39-59, employed at ten California companies. The men were given medical examinations for CHD and a standardized, structured interview to determine their behavior type. Behavior type was determined by reviewing videotapes of the interviews. The 2,648 participants judged to be free of CHD at baseline were followed-up with annual physical examinations to detect new CHD cases during the subsequent 8-1/2 years. The (actual) results of the study are shown in the following diagram and are tabulated in Table 1.

Figure 1
Western Collaborative Group Study 8-1/2 year follow-up

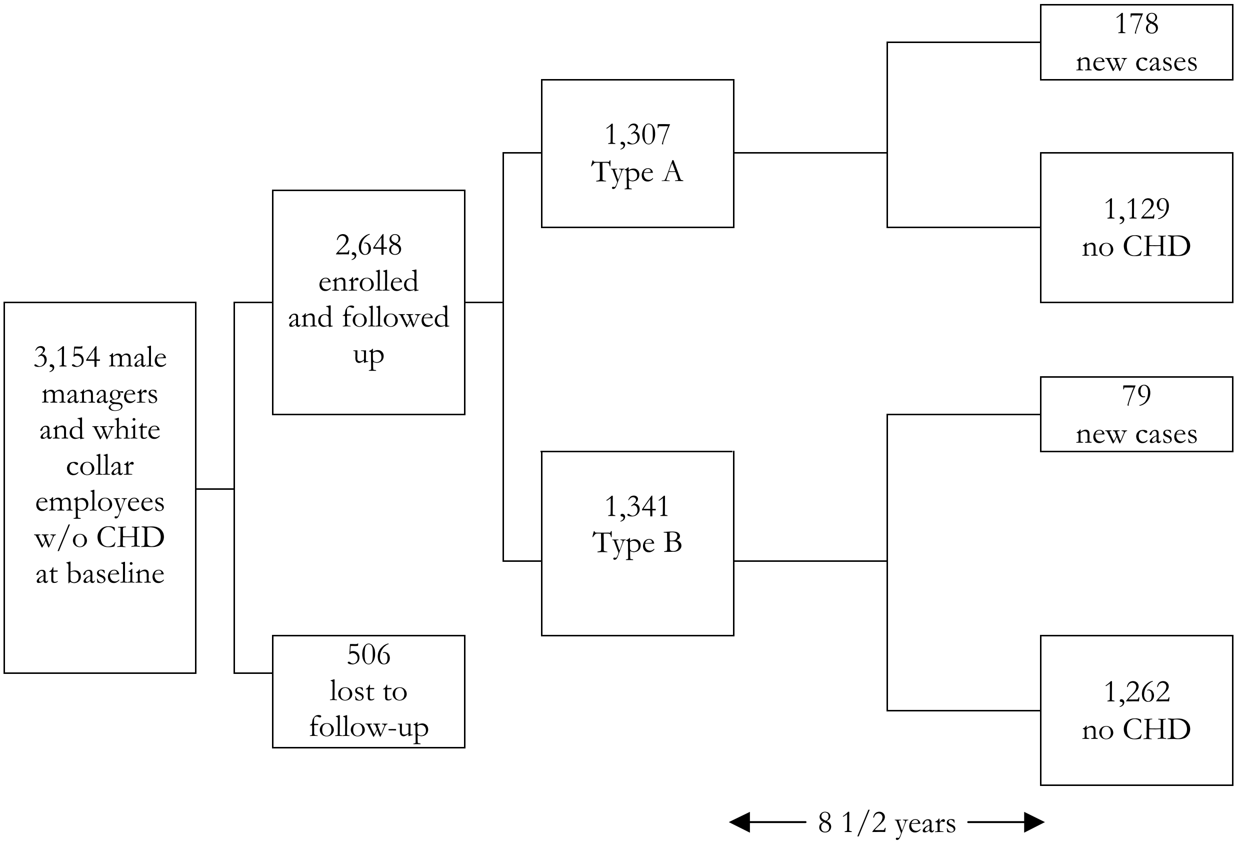


Table 1
Relationship of CHD to Behavior Pattern

	Behavior pattern		
	A	B	Total
CHD cases	178	79	257
No manifest CHD	1,129	1,262	2,391
Total	1,307	1,341	2,648

Since these data come from a cohort study, we would analyze them by estimating the incidence of CHD for the Type A and Type B groups. Even though the risk period for CHD extends beyond the period of observation, we will use cumulative incidence (CI) for simplicity. In these data, the CI is 178/1307 = 0.14 for the Type A group, and 79/1341 = 0.06 for the Type B group. The relative risk (risk ratio, cumulative incidence ratio) is therefore 0.14/0.06 = 2.3

Questions to ask:

There are many aspects of the design and conduct of this study that we would want to inquire about. For example:

> What were the criteria for classifying participants as Type A or Type B?
>
> How many participants were lost to follow-up?
>
> How was CHD defined and diagnosed?
>
> Were the physicians who determined whether a participant was a new case or not aware of the participant's behavior type?

But since our topic today is confounding, we are most interested in the question:

> Do the Type A and Type B groups differ in other factors that might have affected their observed CHD rates?

or, equivalently,

> Are there factors other than behavior pattern that may have been responsible for the observed rates?

(It might be interjected here that the same question would be relevant whether a difference between Type A and Type B had been observed or not).

Hypothetical scenario

Probably most of you know that in the Western Collaborative Group Study, no other factors seemed to explain the difference in CHD incidence between Type A and Type B groups. So here we will depart from the actual study in order to create a scenario in which the difference in the observed incidence for Type A and Type B participants is actually due to differences in cigarette smoking.

Suppose we had obtained the data in the Table 1. How could we see whether the difference in incidence between Type A and Type B groups should be attributed to differences in smoking rather than to behavior type? The traditional and most common approach to answering this question is to break down or stratify the data by cigarette smoking status of the participants. Table 2 shows the results of such a stratified analysis (with hypothetical data).

Table 2
Relationship of CHD to Behavior Pattern,
Stratified Analysis Controlling for Smoking Status [HYPOTHETICAL DATA]

	Smokers		Nonsmokers	
	Type A	Type B	Type A	Type B
CHD	168	34	10	45
$\overline{\text{CHD}}$	880	177	249	1,085
Total	1,048	211	259	1,130

This table shows the relationship between behavior type and CHD, stratified by smoking experience. Now we can compute the (cumulative) incidence of CHD among Type A nonsmokers and compare that to Type B nonsmokers, which will tell us the effect of behavior type when smoking could not possibly account for the results (not counting environmental tobacco smoke). We can also look at the incidence for Type A smokers and Type B smokers, where again we have (to some extent) created groups that are more comparable.

What do we see when we do these calculations? The incidence of CHD among Type A nonsmokers is 10/259 = 0.04, exactly the same as that among Type B nonsmokers (45/1130 = 0.04). We are therefore led to the conclusion that at least among nonsmokers, behavior pattern made no difference. Similarly, the cumulative incidence is the same (0.16) for Type A smokers and Type B smokers. Again, behavior pattern made no difference. Smoking, apparently, made a big difference. This key "extraneous" variable was apparently very unevenly distributed between the two behavior pattern groups and led to our observing a difference we nearly attributed to behavior pattern.

Confounding – a discrepancy between the crude and the controlled

This example illustrates confounding. In the uncontrolled or "crude" table, we saw an association (CIR of 2.3). When we controlled for smoking (which we will assume for the present is the only relevant extraneous variable), we find that there was no association (CIR of 1.0) between our study factor (behavior pattern) and the outcome (CHD). This discrepancy between the crude CIR (2.3) and the stratum specific CIR's (1.0) indicates that there is confounding by smoking status. Stratification is one method of controlling for the confounding effect of smoking. [Please let me emphasize here that the above example is **not** true to life. In the actual study by Friedman and Rosenman, Type A behavior was found to be associated with CHD even when the effects of smoking and other known CHD risk factors were controlled.] It may also be worthwhile to mention that confounding could also happen in the reverse manner, that is, we might see no association in the crude analysis but find that there is one when we stratify. So <u>confounding can create an apparent association or mask a real one</u>.

Confounding arises from unequal distribution of a risk factor

How can the phenomenon of confounding occur? As indicated above, the conditions needed to create confounding (in this rather simplified situation) are that a true risk factor for the health outcome is unevenly distributed between the groups being compared. To see this in the above example, I have rearranged the columns from Table 2. This rearrangement emphasizes that most of the Type A's were smokers and most of the Type B's were not.

Table 3
Relationship between CHD, Behavior Pattern, and Smoking Status
[HYPOTHETICAL DATA]

	Type A behavior pattern			Type B behavior pattern			Both
	Smokers	Non smokers	**Total**	Smokers	Non smokers	**Total**	Grand total
CHD	168	10	**178**	34	45	**79**	257
$\overline{\text{CHD}}$	880	249	**1,129**	177	1,085	**1,262**	2,391
Total	1,048	259	**1,307**	211	1,130	**1,341**	2,648

Although this table was created by rearranging columns in Table 2, it may be more revealing to think of it as providing the underlying story for the uncontrolled (crude) data in Table 1. Notice that Table 1 is contained in this table as the marginals for each of the two subtables (the bolded columns). The subtables show the composition of the Type A group and the Type B group. Clearly, the overwhelming majority (1048/1307 = 80%) of the Type A participants are smokers, whereas the overwhelming majority (1130/1341 = 84%) of the Type B participants are nonsmokers. With such a marked imbalance, it should not be surprising that a risk factor such as smoking could distort the overall (uncontrolled) association. The attributes of a confounder, then, are that it is an independent risk factor for the outcome and is associated with the study factor.

Confounding – misattribution of an observed association

The excess of cases in the Type A group is due, clearly, to the greater proportion of smokers in the Type A group than in the Type B groups. Were we to have gone with the crude value, we would have misattributed the observed difference between groups to behavior pattern rather than to smoking. Confounding can be defined as a distortion in the measure of association due to the unequal distribution of a determinant of the outcome.

Note, however, that the crude association is still "real". The type A participants *did* have a greater incidence of CHD. Confounding arises when we *attribute* that elevated incidence to their being type A, since the higher incidence is really due to their smoking (in this example). But the type A men as a group did indeed have higher CHD incidence. There are situations where the crude association remains important to consider.

Another perspective – weighted averages

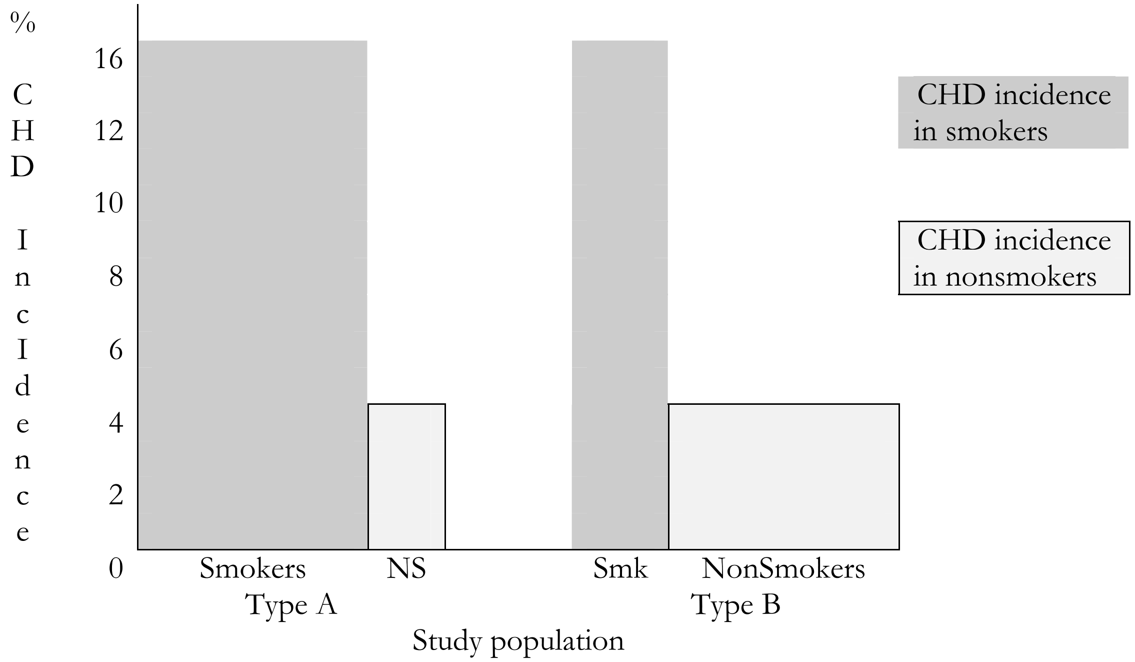
A summary table highlights the incidences and makes the pattern very evident.

Table 4
Incidence of CHD by Behavior Type and Smoking Status
[HYPOTHETICAL DATA]

	Smoking status				
Behavior pattern	Smoker	Nonsmoker	Total		
Type A	0.16	0.04	0.14	←	(incidences
Type B	0.16	0.04	0.06	←	from table 1)
Total	0.16	0.04			

Here it is very clear that when we hold smoking constant (i.e., look down either of the first two columns of incidences), there is no effect of behavior type. When we hold behavior type constant (i.e., look across either of the first two rows), we see that smoking is associated with a fourfold increase in incidence. The marginals of the table are, in effect, weighted averages of the incidences in the interior of the table. The incidences in the bottom row are the same as in the interior of the table – they have to be, because a weighted average of two identical numbers is always that number. The incidences in the rightmost column, however, could be almost any numbers between 0.16 and 0.04 – depending upon the weighting used in averaging 0.16 and 0.04. These concepts can be shown graphically.

CHD Incidence by Behavior Pattern and Smoking Status
[HYPOTHETICAL]



As the diagram shows, the study population can be viewed as consisting of four distinct subgroups, each with a different combination of behavior type and smoking status. If these were the only relevant subgroups, then the incidence rates for each would represent the irreducible "true" state in the study population. The rate for the study population as a whole and for any group in it, e.g., all Type A's, may be regarded as a <u>weighted average of the incidences in the component subgroups</u>, where the weights are the proportional sizes of the component subgroups. Thus the rate in the Type A's is:

$$0.14 = \frac{178}{1,307} = \frac{1,048}{1,307} \times \frac{168}{1,048} + \frac{259}{1,307} \times \frac{10}{259}$$

or symbolically,

$$CI_{CHD|A} = P_{S|A} \times CI_{SA} + P_{\overline{S}|A} \times CI_{\overline{S}A}$$

where:

CI is (cumulative) incidence
P is prevalence or proportion

S indicates smokers ($\overline{S}$ indicates nonsmoker)
A indicates behavior Type A
and the notation S|A means "smokers among (or given) Type A behavior".

Confounding – comparison of weighted averages using different weights

The incidence for any group (e.g., Type A's) can vary from the lowest incidence of any of its subgroups (e.g., nonsmoker Type A's) to the highest incidence of any subgroup (e.g., smoker Type A's). Where in this range the overall group's incidence falls is determined by the size of each subgroup (Type A smokers, Type A nonsmokers) as a proportion of the overall group (all Type A's). Confounding can result when these proportions differ for groups that are being compared.

Since there are many possible ways in which these proportions can differ, confounding can cause an overall (crude) measure of association to overstate, understate, completely obscure, or even invert the association that would be seen in comparisons carried out within the subgroups. As a familiar example, if two populations have different age distributions, then a comparison of their overall (crude) death rates can overstate or understate the picture seen by comparing within specific age groups, even to the point that the comparison of crude rates appears to favor the population that has higher (worse) death rates within each age stratum. Age standardization is a special case of the more general strategy called stratified analysis, which is one primary recourse for controlling confounding.

The limits to confounding

There are limits on the strength of the (secondary) association that can be produced by confounding. For example, given the data in Table 1, a strong effect for smoking and a striking imbalance between the two behavior type groups was required in order for smoking to account completely for the apparent effect of Type A behavior. That is one of the reasons why strength of association is a criterion for causal inference. The stronger the observed association between the disease and the study factor, the less likely that some completely extraneous factor could account for all of the observed association.

Case-control studies

So far in our discussion we have confined ourselves to cohort-type studies. When we turn to the issue of confounding in case-control studies, there are some additional complexities as a consequence of the way in which the base population is represented in the study population. To understand the characteristics of confounding in a case-control study, let us generate such a study from the cohort we considered earlier.

The original cohort consisted of 2,648 individuals with complete follow-up and yielded 257 cases. Ideally, our case-control study would detect all incident cases and would sample from non-cases as the cases occurred (called "density sampling"). To simplify our illustration, however, let us sample our controls from those individuals who were free from CHD at the end of the follow-up period. The following table shows the same cases, with the distribution of controls expected from obtaining a representative sample from the noncases, of size twice the number of cases (i.e., assume 514 controls with the same proportion of Type A's and smokers as found in all noncases in the cohort study). (The numbers in the "No CHD" row are obtained by multiplying the "No CHD" row in Table 1 (i.e., all the noncases) by 514/2391 (0.21) so that the 2,391 noncases become 514 controls.

In this way, the 1,129 Type A's without manifest CHD become 243 Type A controls, and the 1,262 Type B's without manifest CHD become 271 Type B controls.)

Table 5
Expected Results from Case-Control Study [HYPOTHETICAL]

	Behavior pattern			
	Type A	Type B	Total	
CHD cases	178	79	257	
No manifest CHD	243	271	514	← This row is simply 0.21 times the corresponding row in Table 1.
Total	421	350	771	

The odds ratio for this table is [2.5], slightly larger than the risk ratio in the cohort study. [The difference between the odds ratio and risk ratio reflects the CHD incidence in the cohort – the smaller the incidence, the closer the odds ratio would be to the risk ratio.]

Now let us generate, in the same manner, the expected table for smoking and behavior pattern in a stratified analysis:

Table 6
Expected Results for Case-Control Study, Stratified by Smoking Status [HYPOTHETICAL]

	Smokers		Nonsmokers		
	Type A	Type B	Type A	Type B	
CHD	168	34	10	45	
$\overline{\text{CHD}}$	189	38	54	233	← This row is simply 0.21 times the corresponding row in Table 2.
Total	357	72	64	278	

The odds ratios for each table are 1.0, so confounding is again present. Here again we see that the confounding factor is associated with the outcome: the odds ratio for smoking and CHD in the Type B group is 4.6. We also find that smoking is associated with behavior type: the proportion of smokers among Type A noncases is 0.78 whereas among the Type B noncases it is only 0.14 [verify these numbers].

The reason for the above emphasis on ***conditional*** associations ("in the Type B group", "among noncases") rather than ***unconditional*** or crude associations is that a confounding variable must be associated with the exposure under study in the population from which the cases arise (see Rothman and Greenland). It is the control group that provides the estimate of exposure prevalence in the source population. Also, in a case-control study, the totals for different exposure groups (e.g., total

Type A smokers) are not very meaningful quantities, at least for comparison purposes. The reason is that the relationships among these totals largely reflect the (arbitrary) ratio of cases to controls. So the association of exposure that is relevant for confounding in a case-control study is the association between exposure and the potential confounder among the controls.

The reason for not looking within the Type A group is that an association in this group could reflect effect modification between the exposure (Type A behavior) and the covariable, rather than confounding as such. We will elaborate on this matter when we take up effect modification, in the next chapter.

Confounding – a characteristic of the study base

We have said that confounding requires two associations: (1) the confounder must be a risk factor for the outcome or its detection and (2) the confounder must be associated with the exposure. The latter association must exist within the study base (see Rothman and Greenland). This point merits elaboration.

Follow-up study

In a follow-up study, the study base, from which the cases arise, is simply the population being followed, the study population. For confounding to occur, the exposure and potential confounder must be associated in this population. Randomized assignment of an intervention tends to distribute potential confounders evenly across intervention and control groups. To the extent that randomized assignment succeeds, i.e., no extraneous variables will be associated with the intervention, so confounding cannot occur. If, however, the randomization does not "work" so that an imbalance exists for a particular potential confounder, then confounding with respect to that potential confounder can occur. The greater the number of participants, the less likely that any meaningful imbalance will occur by chance.

Case-control studies

In a case-control study, the study base is the underlying population that is being followed through the window of the case-control design. For confounding to occur, the exposure and potential confounder (risk factor) must be associated in that underlying population (source population from which cases arise). But since the investigator observes that population only indirectly, the matter is trickier. However, if there is no association between the potential confounder and exposure in the study base, then confounding does not occur even if we do find the potential confounder and exposure to be associated within the control group of our case-control study (Miettinen and Cook, cited in Rothman, page 93).

This somewhat surprising result is easily illustrated. Suppose we are observing a population over time to examine an association between a suspected occupational carcinogen and a cancer that is also strongly (IDR=10) related to cigarette smoking. Suppose also that the occupational exposure is in fact a carcinogen and that in this population smoking is not associated with the occupational exposure. If we assume a baseline rate of 3 cases/1,000 person-years and an IDR of 3.3 for the occupational carcinogen, the follow-up of the population might produce the following table.

Incidence rates, population sizes, and number of cases for hypothetical data on an occupational exposure and smoking

	Smokers		Nonsmokers	
	Exposed	Unexposed	Exposed	Unexposed
	(1)	(2)	(3)	(4)
1. Number of cases	300	90	70	21
2. Population size (person-years)	3,000	3,000	7,000	7,000
3. Incidence density per 1,000 py	100	30	10	3
	IDR = 3.3		IDR = 3.3	

With a hypothetical 7,000 person-years of observation for nonsmokers who are also not exposed to the carcinogen, the assumed baseline incidence rate of 3/1,000 py will produce an expected 21 incident cases. If the amount of person-time among exposed nonsmokers is also 7,000 py, then we would expect 3.3 × 3/1,000 py × 7,000 py ≈ 21 cases for that group. If person time for exposed and unexposed smokers is 3,000 py for each group, then we expect 300 and 90 incident cases, respectively, if the IDR for the occupational exposure is the same among nonsmokers and smokers and the IDR for smoking is 10, regardless of occupational exposure.

Note that this hypothetical population has been constructed so that the proportion of exposed person-years is 50% among smokers (columns 1 and 2), among nonsmokers (columns 3 and 4), and overall, i.e., no association between smoking and the occupational exposure. Similarly, the proportions of person-years for smokers among exposed (columns 1 and 3) and unexposed (columns 2 and 4) are each 30% (3,000/[7,000+3,000]). The crude IDR for the occupational carcinogen is therefore 3.3 (be certain that you can derive this IDR), which is identical to the IDR for the exposure among smokers and among nonsmokers. Thus, confounding is not present.

Suppose now that we were to conduct a case-control study in this population during the same period of time. If there is a cancer registry we might hope to identify and include all 481 cases (see row 1 in the following table, which is identical to row 1 in the preceding table). If we obtain a 5% representative sample of the population as our control group, then the distribution of smoking and the occupational carcinogen in our control group (row 2 in the following table) will be the same as the distribution of these variables in the population-time in row 2 of the preceding table (30% smokers and 50% exposed to the occupational carcinogen, with no association between these two). The OR (be certain that you can calculate this) will be identical to the IDR of 3.3, above. In this case-control study with an (**unbiased**) control group that is directly proportional to the study base, there is no confounding.

Different control groups for hypothetical case-control study of an occupational exposure and smoking

	Smokers		Nonsmokers	
Row	Exposed	Unexposed	Exposed	Unexposed
#	(1)	(2)	(3)	(4)
1. Number of cases	300	90	70	21
2. Proportional controls	150	150	350	350
	(OR = 3.3)		(OR = 3.3)	
3. Biased controls	250	150	250	350
	(OR = 2.0)		(OR = 4.7)	

Suppose, however, that controls are selected in a biased fashion, producing a **biased control group** (row 3 in the second table) in which smoking and exposure **are** associated (verify this fact; try, for example, computing the OR for smoking in relation to exposure). Reflecting the biased control group, the stratum-specific IDR's are no longer 3.3. However, in this chapter our focus is the **crude association** and whether it accurately represents the true situation (which in this instance we constructed, rather than having to regard the stratified associations as the true situation). The crude OR from the above table, using the cases on row 1 and controls from row 3, is (do try computing this before reading the answer) (370 × 500) / (111 × 500) = 3.3.

Thus, even with this biased control group the crude OR remains unconfounded. Yet, the potential confounder (smoking, a causal risk factor for the outcome) is indeed associated with the exposure in the (biased) controls. [Several ways to see this association are:

The odds of exposure among smokers (cols. 1 and 2) are 250/150, quite different from the odds of exposure among nonsmokers (cols. 3 and 4: 250/350), producing an odds ratio between smoking and exposure of OR = 2.3).

Proportionately more smokers are exposed [250/(250 + 150) = 0.63] than are nonsmokers [250/(250 + 350) = 0.42].

The odds of smoking among exposed (cols. 1 and 3) are 250/250, quite different from the odds of smoking among the unexposed (cols. 2 and 4): 150/350), producing, of course, the same odds ratio, 2.3).

Proportionately more exposed are smokers [250/(250+250) = 0.5] than are unexposed [150/(150 + 350) = 0.3].

The potential confounder, smoking, is also associated with the outcome in the unexposed (e.g., IDR = 30 per 1,000py / 3 per 1,000py in the study base, OR = (90 × 350) / (21 × 150) in the case-control study with either control group. <u>Thus, it is possible to have a risk factor that is associated with exposure in the noncases yet not have confounding</u>.

Further insight can be gained by considering the mechanism that causes confounding, as illustrated in the Type A behavior example. Confounding results from an imbalance between exposed and unexposed groups in regard to a disease determinant. If the potential confounder increases disease risk and the potential confounder is associated with the exposure, then incidence of disease in the exposed will be boosted relative to that in the unexposed, due to the confounder. This <u>disproportionate increase in incidence, and therefore in cases, will increase the odds of exposure in a representative group of cases</u>. If the confounder is not controlled in the analysis, this increased odds will cause confounding of the exposure-disease association.

The (exposure) OR for the outcome is simply the ratio of the exposure odds in the case group divided by the exposure odds in the control group. The exposure odds in the case group is obviously not affected by anything that happens to the control group (including matching, incidentally). So a distortion of the crude OR will have to come from a change in the exposure odds in the control group. So long as the bias in the control group does not cause its **crude** exposure odds to differ from those in the source population (e.g., 0.5/0.5=1.0 in our occupational carcinogen example), the crude OR will remain the same as in the source population, i.e., unconfounded.

In most case-control studies we have little independent information about the study base, so the control group provides our window into the study base. If the control group is biased, then our view of the study base is distorted, and we may conclude that the condition for confounding (i.e., a risk factor for the disease is associated with the exposure in the noncases) is met. Due to such a biased control group, controlling for the potential confounder will introduce bias in this analysis (e.g., in the above example, the stratum-specific OR's are different from the correct value of 3.3). However, a weighted average of stratum-specific OR's may be close to the crude value.

Statistical tests for confounding

Since confounding requires an association between the potential confounder and the exposure, investigators sometimes present statistical tests of the differences in potential confounders between exposure groups. If the groups do not differ significantly, the investigators conclude that confounding will not occur. This practice will often yield a correct conclusion, though it is somewhat off the mark.

Statistical tests of significance address the question of whether or not there is an association between the exposure and the potential confounders <u>beyond that likely to arise by chance alone</u>. But confounding depends upon the <u>magnitude of association</u> (e.g., odds ratio, prevalence ratio), rather than on the strength of evidence that it did not arise by chance. So a large but "nonsignificant" difference can have more potential to cause confounding than a small but "highly significant" difference. The reason for this apparently paradoxical statement is that statistical significance depends upon the <u>magnitude of the number of exposed and unexposed</u> participants, so that nearly any association will be statistically significant if the study is sufficiently large and nonsignificant if it is sufficiently small. The presence or extent of confounding, however, is not affected by scaling up or down the number of participants.

Confounding, then, is a function of the magnitude of associations, rather than of their statistical significance. Since strong associations are likely to be statistically significant, statistical tests comparing exposed and unexposed groups can be a convenient device for identifying associations that may be strong enough to cause confounding, which is why the procedure often yields the correct conclusion about the need to control for confounding. Some (see Rothman and Greenland) have suggested using significance testing with a value for alpha (Type I error probability) of 0.20, to increase the power to detect differences that may be important in regard to confounding. But as a guide to likely confounding, statistical tests are somewhat beside the point. (There is a subtle but valuable distinction to be made between statistical tests to evaluate confounding and statistical tests to assess whether randomized allocation to treatment or control "worked". Since randomized allocation attempts to operationalize "chance", the number and size of observed differences between treatment and control groups should not often exceed what we expect from chance, which is precisely what statistical tests are designed to evaluate. If there are more differences than there "should be", that may indicate some problem in the implementation of the randomization. It would also be expected that control for these differences would be needed.)

Components of the crude relative risk

There are several other aspects of confounding that it will be instructive to consider. The first of these is a method, due to Miettinen (Miettinen OS: Components of the crude risk ratio. *Am J Epidemiol* 1972; 96:168-172) for allocating an observed association to a component due to confounding and a component due to the study factor of interest. According to Miettinen, the crude risk ratio (or odds ratio) may be regarded as the product of a "true" risk ratio and a component due to confounding. In the examples we have considered thus far, the whole of the observed association has been due to confounding, i.e., to the effect of smoking. But it is also possible to have an association that remains, though stronger or weaker, after the effects of a confounder have been removed.

The following hypothetical data illustrate Miettinen's concept. Suppose that you are carrying out a case-control study to investigate whether trihalogenated hydrocarbons that occur in chlorinated drinking water containing organic matter increase colon cancer incidence. You collect data on all cases in a multi-county region during several years and assemble a control group using random-digit dialing. You interview cases and controls about their source of drinking water (treated surface water versus well or bottled water) and, because other studies have suggested that some unknown factor in urban living increases colon cancer incidence, you also collect data on urban-rural residence. The crude analysis of your data yields the following table:

Table 7a
Colon cancer and drinking water (hypothetical case-control data)

	E	$\bar{E}$	Total
Colon cancer cases	170	80	250
Controls	80	170	250
Total	250	250	500

The crude OR for this table is (170 x 170) / (80 x 80) = 4.5. Is it confounded by rural-urban residence?

We can investigate confounding by stratifying the data by urban-rural residence and examining the stratum-specific OR's:

Table 7b
Colon cancer and drinking water (hypothetical case-control data)

	Rural			Urban			Crude	
	E	$\bar{E}$		E	$\bar{E}$		E	$\bar{E}$
D	20	30	D	150	50	D	170	80
$\bar{D}$	50	150	$\bar{D}$	30	20	$\bar{D}$	80	170

The OR's in both the rural and urban strata are 2.0, so we know that the crude OR is confounded – it overstates the "true" OR, making a moderate association appear as a strong one. How much of the crude OR can be attributed to confounding? Miettinen suggests that the OR due to confounding is the OR for the association that would be observed even if the exposure (trihalogenated hydrocarbons) had no effect on the outcome (colon cancer). If the exposure has no effect on the outcome, then whatever association remains in the crude analysis must be due entirely to confounding.

So to obtain the OR attributable to confounding, we can eliminate the true association between trihalogenated hydrocarbons and colon cancer. In the above example, we regard the stratum-specific tables as displaying the true relationship (i.e., we are assuming that there is no selection bias, or information bias and that the only potential confounder is rural-urban residence as a dichotomous variable measured without error). So we will "eliminate" the true association from the stratum-specific tables. Then we can combine the modified stratum-specific tables into a new crude table and compute a new crude OR. That OR must entirely reflect confounding, because the true association no longer exists.

Since the OR is the crossproduct ratio for the four cells of a table, we can change the OR by changing any cell of the table. By convention, we change the "a"-cell (exposed cases) to what it would contain if there were no association between the study factor and the disease. Here, if the D,E cell in the rural stratum contained 10 instead of 20, then the OR for the rural stratum would be 1.0, i.e., no association. Similarly, if the D,E cell in the Urban stratum contained ___ (your guess?) instead of 150, then the OR for that stratum would likewise be 1.0. The revised tables are shown below:

Table 7c
Modified tables for Colon cancer and drinking water

	Rural			Urban			Modified crude			Original crude	
	E	$\bar{E}$		E	$\bar{E}$		E	$\bar{E}$		E	$\bar{E}$
D	10	30	D	___	50	D	85	80	D	170	80
$\bar{D}$	50	150	$\bar{D}$	30	20	$\bar{D}$	80	170	$\bar{D}$	80	170

The OR for the modified crude table, and therefore the component attributable to confounding, is 2.25. Interestingly, this figure is the same as the quotient of the original crude (4.5) and controlled odds ratios (2.0). Indeed, this relationship holds in general: the crude OR equals the product of the controlled OR and the component attributable to confounding:

$$\text{Crude odds (or risk) ratio} = \text{Component due to study factor} \times \text{Component due to confounding}$$

So the component (of the crude ratio) attributable to confounding is the degree of association "expected" from the distribution of the potential confounder (in this case, rural-urban residence), i.e., from the fact that the potential confounder is distributed differently in exposed and unexposed persons in the study base.

Another way to look at this relationship is that the component attributable to the effect of the study factor, i.e., the unconfounded association, can be written as:

$$\text{Component due to study factor} = \frac{\text{Crude odds (or risk) ratio}}{\text{Component due to confounding}}$$

So the component (of the crude ratio) attributable to the study factor, i.e., the unconfounded, or "true", association, can be regarded as the ratio of an "observed" association to an "expected" association. Expressing the relationship in this way is reminiscent of the standardized mortality ratio (SMR), which is also a ratio of an "expected" to an "observed". In fact, Miettinen refers to the controlled OR above (i.e., the component due to the study factor) as an "internally standardized odds ratio", which is simply the odds ratio version of the SMR. It is also interesting to note that the stratum-specific OR's are also ratios of "observed" to "expected", in that these OR's are equal to the ratio of the observed number of exposed cases (the contents of the "a-cell") to the expected number in the absence of a true association.

By this point you may well be wondering how much of this you need to know to practice epidemiology or control for confounding. The answer is that this particular formulation is not essential, but seeing confounding from this perspective is another aid to understanding the closely interrelated concepts of confounding, stratified analysis, standardization, and even the counterfactual framework of causal inference. If the true causal comparison is between the experience in an exposed group and what their experience would have been in the absence of exposure, the SMR might be regarded as the most relevant adjusted measure of association, since it is the ratio of the observed rate in the exposed group to the rate that would be expected for them if they were not exposed (assuming that the rates in the study population differ from those in the standard population due only to the standardizing factor and the exposure).

Matched studies

Since confounding is a problem of comparison, a principal aim of study design is to obtain groups that are comparable with regard to determinants of the outcome. In experimental designs, this aim is perhaps the principal motivation for randomized assignment of the study factor. Since randomized allocation does not guarantee the equal distribution of all relevant factors (though in very large studies the probability of equal distribution is very high), ***prestratification*** (also called "***blocking***") may be employed to enforce identical distributions when sample size is small. Prestratification involves first placing participants into groups according to their configuration of risk factors and then performing separate randomizations within each group. The procedure generally increases statistical efficiency (degree of precision per trial participant) (see Rothman and Greenland, p161).

Follow-up studies

In a nonrandomized study, where the investigator does not have the opportunity to assign the study factor, the analogous procedure to prestratification is ***matching***. In matching, the participants in the comparison group (i.e., the unexposed group in a follow-up study or the control group in a case-control study) are selected so as to resemble the ***index group*** (the exposed in a follow-up study or the cases in a case-control study) on one or more relevant factors. When the unexposed group in a follow-up study has been matched to the exposed group on all relevant factors, so that the two groups differ only in terms of exposure to the study factor of interest, then the incidences in the two groups can be compared with no danger of confounding by the matching variables. In practice, however, competing risks and/or loss to follow-up can introduce differences. For this and other reasons (see Rothman and Greenland, p160), matched cohort studies are not common.

In any case, neither prestratification nor matching is required to avoid confounding, since confounding can be controlled in the analysis of the study results – providing there is adequate overlap in risk factor distributions between groups. For this reason, the primary purpose of matching is to increase statistical efficiency by ensuring sufficient overlap (which therefore indirectly aids in controlling confounding).

Case-control study

In a case-control study the situation is, as usual, not as straightforward. Because of the nature of the case-control study design, matching does not avoid confounding by the matching factor(s). Moreover, by changing the composition of the control group, matching in a case-control study can even cause the crude (uncontrolled) analysis to be biased. How can this be?

Since a case-control study selects participants according to disease, matching means ensuring that the case and control groups are the same in respect to the potential confounders. However, as we saw earlier, confounding depends on the comparability of exposed and unexposed groups in the study base, not between cases and controls in the study population. Although ensuring that cases and controls are similar with respect to potential confounders may facilitate control for confounding (through greater statistical efficiency), matching controls to cases does not change the study base and thus cannot alter the exposure odds among cases. But confounding arises because the exposure odds in cases is influenced by a population imbalance in a cause of the outcome.

Furthermore, by selecting the control group in a way that makes it conform to the case group in age, sex, treatment facility, or other factors, the investigator can cause the overall control group to have a different prevalence of exposure than that in the study base, which the control group seeks to reflect. Of course, a matched control group can still provide a correct estimate of exposure prevalence within each configuration of risk factors. So there need be no problem as long as the analysis takes account of the matching. If the matched analysis and unmatched analysis yield the same results, then the unmatched analysis can be used, and for simplicity often is unless the matched analysis provides greater precision.

Example of matching in a case-control study

The following example may help to clarify these concepts. Consider another study of colon cancer and drinking water, presented in the following table. This time the stratum-specific population sizes and prevalences of exposure to chlorinated drinking water are presented, along with the number of cases and the prevalence of exposure among cases.

Colon cancer and drinking water (hypothetical data)

Residence	Population size	% of total pop. with chlorinated drinking water	# of colon cancer cases	% of cases with chlorinated water
Rural	400,000	20 %	30	40 %
Urban	600,000	80 %	90	90 %
Total	1,000,000	56 %	120	___ %

An investigator conducting a case-control study in this population and selecting community controls without matching, would observe an exposure prevalence of 56% (i.e., an average of the urban- and rural-specific exposure prevalences, weighted by their respective population sizes: [0.20(400/1000) + 0.80(600/1000)]). In contrast, a control group matched to the distribution of cases would have an exposure prevalence of 65% [0.20(30/120) + 0.80(90/120)], since in this case the two prevalences are weighted by the proportions of rural and urban cases, rather than the proportions of rural and urban residents in the population.

The prevalence of exposure in the matched control group, 65%, is a distorted estimate of the overall prevalence of exposure in the population as a whole. But the estimate is not a problem when our analysis takes rural-urban residence into account, since the stratum-specific exposure prevalences are still correct and we know the proportions of rural and urban residents in the population. If the exposure prevalence (right-most column) is 40% in rural cases and 90% in urban cases, then the odds ratios are 2.67 (rural) and 2.25 (urban), 2.70 (crude, unmatched controls) and 1.85 (crude, matched controls). Thus neither the matched nor the unmatched controls give a correct OR for a crude analysis. In contrast, a stratified analysis that takes residence into account will yield a valid odds ratio estimate with either control group. (Suggestion: derive all of these OR's.)

For a fuller treatment of matching, see chapter 10 of Rothman and Greenland. According to these authors, though there are circumstances where it is clearly desirable or not desirable, the value of matching in case-control studies is a complex question.

Potential confounders versus actual confounders

An issue of considerable practical and theoretical importance is how to choose which variables to investigate as confounders? As we saw above, to be a confounder a variable must be associated with both the disease and the exposure. Thus when through matching in a cohort study we ensure that the distribution of potential confounders is identical in both exposure groups (i.e., there is no association between these variables and exposure), then the former cannot confound our results (assuming no bias from competing causes of death and other attrition mechanisms). Apart from that situation, we must control for potential confounders in the analysis of the study to see whether or not they have distorted the observed association (which implies that we have remembered to measure them!).

Investigation of whether a variable is a potential confounder or an actual confounder is thus generally a matter of empirical determination in our data. In practice, therefore, the question of whether or not variable X is a confounder is a side issue. Our primary concern is to obtain a valid estimate of the relationship between study factor and outcome. If we have to control we do; if we do not need to, we may not. In either case we are not particularly concerned, ordinarily, about concluding that such-and-such a variable is a confounder.

But which variables to regard as potential confounders, i.e., which variables must be measured and possibly controlled in order to obtain a valid estimate of the association between study factor and outcome, is a matter of first importance. Our decisions here depend upon our understanding of which variables other than our study factor might explain or account for an observed relationship (or lack thereof). Thus, the decision about whether a variable ought to be considered for control as a potential confounder rests first and foremost on our conceptual model.

First and foremost, a potential confounder must have some relationship to the occurrence of the disease or other outcome. The potential confounder must increase the probability that the disease will occur or must shorten the time until the disease occurs. If not, why should we attribute an observed association to that variable rather than to our study factor? (Since disease occurrence must be observed, a factor that affects disease detection may also qualify.) Furthermore, if the relevant variable occupies an intermediate position in the hypothesized causal chain linking the study factor to the disease, then again, how could that variable rather than the study factor be the "true" cause of an observed association? (If I persuade George to rob a bank and the police find out, can I persuade the judge to set me free because apart from what George did I did not rob anything?) Thus, in stratifying on smoking status in our Type A - CHD example, we are assuming that the association between Type A behavior and smoking arises due to a common antecedant cause (e.g., inadequate coping skills in a high-pressure occupational environment) or due to an effect of smoking status on behavior pattern, but not due to an effect of behavior pattern on smoking status, which would make smoking an intervening variable and therefore not appropriate for control in this way (Kaufman and Kaufman, 2001).

In practice, however, it is often difficult to make definite decisions about which variables are true risk factors, which are intervening variables, and so on, so that a cautious approach is to obtain data on as many potentially relevant variables as possible, explore the effects of controlling them in the analysis of the study, and then attempt to make sense out of the results. Consider, for example, a study of the effect of overweight on CHD incidence. Since overweight increases cholesterol and blood pressure levels, both of which are causal risk factors for CHD, then the crude association between overweight and CHD will reflect some combination of:

1. a direct effect of overweight on CHD if such exists,
2. an indirect effect of overweight on CHD due to the effect of overweight on cholesterol and blood pressure, which in turn increase CHD risk
3. possible confounding, if cholesterol and blood pressure are higher in people who are overweight not because of an effect of overweight but due to some other reason (e.g., diet, sedentary lifestyle, genetic factors).

Should we control for blood pressure and cholesterol when estimating the association between overweight and CHD? If we do not, then our measure of association will be distorted to the extent that confounding is present. If we do control by the usual methods, however, our measure of association will be distorted to the extent that overweight has its effects on CHD through increases on blood pressure and cholesterol.

For another example, consider the problem of studying whether sexually transmitted diseases such as gonorrhea increase the risk of acquiring HIV and whether condom use decreases the risk. Should the relationship between STD and HIV seroconversion be controlled for condom use? Should the relationship between condom use and HIV incidence be controlled for STD? Both condoms and STD appear to affect the risk of acquiring HIV infection, but condoms are also a means of preventing STD, which in that sense can be regarded as a variable located on the causal pathway from condoms to HIV. 'Furthermore, an obligatory causal factor for both sexually-acquired STD and/or HIV is sexual contact with an infected partner. "Risky" partners have a higher probability of being infected, and the more of them, the greater the risk of exposure to the infection. Should we control for the number of risky partners in investigating the relationship among condoms, STD, and HIV? But risky partners are also a risk factor for STD, so that STD can be regarded as an intermediary variable between sex with risky partners and HIV. Thus, thinking through which variables to control for in a web of causation can itself be confounding! Greenland, Pearl, and Robins (1999) present a system of causal diagrams for describing and analyzing causal pathways to identify what variables must be controlled. Among other points, they explain that controlling for a variable can in some situations create confounding which would otherwise not occur. In general, control for confounding (and interpretation of data in general) is founded on assumptions of causal relationships involving measured and unmeasured variables. Data alone are inadequate to resolve questions of causation without these assumptions. Methodological understanding in this area is expanding (see the articles by Greenland, Pearl and Robins and Kaufman and Kaufman). However, limited knowledge of causal relationships in addition to the one under study and the likely existence of unmeasured but important variables will remain fundamental stumbling blocks for observational research..

Controlling for sociodemographic variables

Nearly all epidemiologic investigations control in some way or other for sociodemographic variables (e.g., age, gender, race, socioeconomic status). As we saw in the chapter on Standardization, comparisons that do not control for such variables can be very misleading. However, there are significant issues of interpretation of adjustment for sociodemographic factors, partly because associations with sociodeomographic factors likely reflect the effects of factors associated with them, and some of these factors may be intervening variables. For example, studies of ethnic health disparities often attempt to control for differences in socioeconomic status. However, it has been argued that socioeconomic status is an intervening variable between ethnicity and health outcomes, so that its control by the usual methods is problematic (Kaufman and Kaufman, 2001). The problem of interpretation is compounded when the persistence of an association with ethnicity, despite control for other factors, prompts the investigator to make an unwarranted inference that a genetic factor must be at work. It is also worth noting that the crude association presents the situation as it exists. Even if the causal explanation indicates other factors as responsible, the fact of disproportionate health status remains an issue to be dealt with. Moreover, a remedy may not require dealing with the "real" cause.

"Collapsibility" versus "comparability"

Although the problem of confounding and the need to control for it have long been a part of epidemiology and other disciplines, theoretical understanding of the confounding has been developed largely since Miettinen's articles in the mid-1970's. Two opposing definitions or perspectives have been debated during that time, one called "comparability" and the other called "collapsibility".

In the ***comparability*** definition, advocated by Sander Greenland, James Robins, Hal Morgenstern, and Charles Poole, among others (see bibliography for article "Identifiability, exchangeability, and epidemiological confounding" and correspondence "RE: Confounding confounding"), confounding is defined in relation to the counterfactual model for causal inference, described in the beginning of this chapter. Confounding results from noncomparability, i.e., a difference between the distribution of outcomes for the unexposed group to what would have been observed in the exposed group if it had not been exposed. Since the latter value is hypothetical and unobservable, the comparability definition cannot be directly applied, though it has some theoretical advantages as well as practical implications.

In the ***collapsibility*** definition, advocated by D.A. Grayson (*Am J Epidemiol* 1987;126:546-53) and others, confounding is present when the crude measure of association differs from the value of that measure when extraneous variables are controlled by stratification, adjustment, or mathematical modeling. If 2 x 2 tables for different strata of a risk factor (as in the Type A example above) produce measures of association (e.g., RR's) that are essentially equivalent to the measure of association for the "collapsed" 2 x 2 table (disregarding the risk factor used for stratification), then there is no confounding in regard to that measure of association. The collapsibility definition is readily applied in practice and is widely used. Disadvantages for this definition are that it makes confounding specific to the measure of association used and the particular variables that are being controlled.

Fortunately for practicing epidemiologists, the two definitions generally agree on the presence or absence of confounding when the measure of effect is a ratio or difference of incidences (proportions or rates). The major practical problem arises when the measure of association is the odds ratio (unless the situation is one where odds ratio closely estimates a risk or rate ratio, e.g., a rare outcome). Further explanation of this and related issues are presented in the Appendix (and in Rothman and Greenland).

Controlling confounding

Now that we are all impressed with the importance and value of taking into account the effects of multiple variables, what are some of the analytic approaches available to us? The principal ones are the following:

- Restriction
- Matching
- Stratified analysis
- Randomization
- Modeling

Restriction

When we adopt restriction as an approach, we are in effect opting not to attempt a multivariable analysis – we simply restrict or confine our study to participants with particular characteristics (e.g., male, age 40-50, nonsmokers of average weight for height, with no known diseases or elevated blood pressure) so that we will not have to be concerned about the effects of different values of those variables. Restriction of some sort is nearly always a part of study design, since virtually all studies deal with a delimited geographical area, specific age range, and so on, though the motive may be feasibility rather than avoidance of confounding. If it is known or suspected that an association is strongest in a particular population subset, then it may make sense to focus studies in that group. Or, if there are few data available that apply to a particular population, it may make sense to restrict study participants to persons in that population. Restriction is also useful when an important variable (particularly a strong suspected risk factor) is very unevenly distributed in the target population, so that it will be difficult and expensive to obtain enough participants at the less common levels of that variable.

Considerations such as these have often been cited as the reason why so many studies in the United States have enrolled only white participants, often only white males. For example, in many potential study populations (defined by geography, employment, or membership), there are (or were) too few members of minority groups to provide sufficient data for reliable estimates. Reasoning that in such a situation stratified analysis by race/ethnicity is essentially equivalent to restriction to whites only, investigators often simply limited data analysis or data collection to whites. The reasoning behind limiting studies to males has been that because of the very different disease rates in men and women, studies of, for example, CHD in middle-aged persons, require many, many more women than men in order to obtain a given number of cases (and therefore a given amount of statistical power or precision). The fact that until about the 1980's the number of women epidemiologists and their representation in policymaking were fairly small may also have had some influence.

The reasons for restricting study participants according to race/ethnicity are more complex. If race/ethnicity (the term "race" virtually defies precise definition) is not a risk factor for the outcome under study, then there is no need to stratify or restrict by race/ethnicity in order to control for confounding. But the United States' ever-present racial divide and its accompanying pervasive discrimination, widespread exploitation, frequent injustices, recurrent atrocities, and continuing neglect by the dominant society have created intellectual, attitudinal, political, and logistical barriers to race-neutral research (see bibliography; the states of the American south maintained legally-enforced apartheid well into the 1960's, and extra-legally enforced apartheid continues to this day). Many of these issues have also arisen for other United States populations with ancestry from continents other than Europe.

The concept of race as a powerful biological variable capable of confounding many exposure-disease associations has its historical roots in 19th century "race science", where various anatomical, physiological, and behavioral characteristics, assumed to be genetically-based, were interpreted as demonstrating the relative superiority/inferiority of population groups and justifying the subordination by whites of colored peoples (Bhopal, Raj. Manuscript in preparation, 1996; see bibliography for additional references). Various diseases and conditions were linked to racial groups (including "drapetomania", the irrational and pathological desire of slaves to run away, and "dysaethesia Aethiopica" ["rascality"]). One reads in medical books from the period that blacks are "an exotic breed".

Most of these ideas are now widely discredited, though by no means extinct (see Carles Muntaner, F. Javier Nieto, Patricia O'Campo "The Bell Curve: on race, social class, and epidemiologic research". *Am J Epidemiol*, September 15, 1996;144(6):531-536). But until recently the vast majority of epidemiologic study populations have been white, English-speaking, urban or suburban, and not poor. A scientific basis linking race itself to health outcomes has emerged for only a handful of conditions (most prominently skin cancer and sickle cell trait and disease). But the suspicion that race could be a risk factor is difficult to dispel, in part because it is reinforced by the many race-related differences in health outcomes. These differences presumably arise from differences in diet and nutrition, physical and social environment, early life experiences, economic resources, health care, neighborhood characteristics, social interactions, experiences of discrimination, lifestyle behaviors, and the host of other factors that affect health and wellbeing, but race is a much more easily measured (if not defined) surrogate for risk.

In addition, many of these differences present logistical challenges (e.g., unfamiliarity of [primarily white, middle class] researchers and staff in studying persons and communities from other backgrounds, distances from research institutions, limited infrastructure, scarcity of questionnaire and other measurement tools that have been validated on multiple racial/ethnic groups, among others). The practical aspects of epidemiologic studies typically demand a great deal of time, effort, and cost, so it is natural to seek ways to reduce these.

Whatever the motivations and their merits, the overall impact of focusing research on white, English-speaking, urban/suburban, nonpoor populations is a scarcity of knowledge, research expertise, data collection tools, and ancillary benefits of participation in research (e.g., access to state-of-the-art treatments, linkages between health care providers and university research centers) for other populations – even for conditions for which these populations have higher rates or for which there are scientific or public health reasons for questioning the applicability of findings for Americans of European extraction to people of color or Latino ethnicity.

Since about the mid-1980's, partly in response to prodding from Congressionally-inspired policies of the National Institutes of Health and Centers for Disease Control and Prevention that now require all grant applicants to provide strong justification for not including significant numbers of women and minorities in proposed studies, research on understudied populations has increased substantially. These policies and the measures taken to enforce them have created new challenges for epidemiologists and in many cases have increased the complexity of epidemiologic studies. However, epidemiology cannot on the one hand claim that it is an essential strategy for improving public health and on the other hand largely ignore one-fourth (minorities) or five-eighths (women plus minority men) of the population.

Several years ago the American College of Epidemiology issued a "Statement of Principles on Epidemiology and Minority Populations" (*Annals of Epidemiology*, November 1995;5:505-508; commentary 503-504; also under "policy statements" on the College's web site, www.acepidemiology.org) recognizing the importance of minority health for public health, of improving epidemiologic data on minority populations, and of increasing ethnic diversity in the epidemiology profession. The Statement has been endorsed by the governing bodies of various epidemiology and public health organizations, including the Council on Epidemiology and Prevention of the American Heart Association, North American Association of Central Cancer Registries, Association of Teachers of Preventive Medicine, American College of Preventive Medicine, American Statistical Association Section on Epidemiology in Statistics, American Public Health Association, and the epidemiology faculties at numerous institutions (e.g., Harvard, UNC, University of Massachusetts at Amherst, and University of Texas Health Sciences Center). In January 2000, the U.S. Department of Health and Human Services announced the goal of eliminating racial/ethnic disparities in health by the year 2010. This challenge and the related one of bringing racial/ethnic diversity to the epidemiology profession are fundamental to public health in the United States at least.

Matching

As discussed earlier, confounding by a risk factor(s) can be avoided in a follow-up study by ensuring, through matching, that the various exposure groups have the same (joint) distributions for those risk factor(s). Thus in a cohort study or an intervention trial, we can select participants at one exposure level and then select participants for another exposure level (including "unexposed") from a larger candidate population according to the distribution of selected risk factors in the first group.

For example, consider a retrospective cohort study to investigate whether players in collegiate revenue sports (e.g., football, basketball), when they reach age 60, are more likely to have altered evoked potentials in response to auditory stimuli, suggestive of differences in neurologic function. The exposed cohort might consist of former basketball players from the team rosters of several universities, the comparison (unexposed) cohort of former students from the same universities during the same years.

Since measurement of evoked potentials is a lengthy and expensive process, we want each participant to be as informative as possible, in order to minimize the total number of participants needed for the study. If we choose unexposed participants completely at random, it is likely that they will differ from the basketball players in a number of ways (measured during their college years) that might affect evoked potentials – height, physical health, strength, agility, coordination, age (for example, the basketball players are unlikely to be mature students returning to complete a degree after taking time off to support a family), parental education, SAT (Scholastic Aptitude Test) scores (because athletes may be recruited for their talent even if their academic records are less competitive), and sex (revenue sports are, or at least were, all male). Some of these characteristics may affect evoked potentials. Thus, comparisons of evoked potentials at age 60 between the basketball players and the other alumni could be confounded by different distributions of these and other variables.

When we attempt to control for these differences, we may find that they are so large that there are basketball players (e.g., those taller than 6-feet) for whom there are very few or no comparison subjects and comparison subjects (e.g., those shorter than 5-feet, 8-inches) for whom there are very few if any basketball players. But strata with few basketball players or with few comparison subjects provide less information for comparing evoked potentials than do strata where the two groups are present in approximately equal numbers. The findings from the analysis that controls for confounding will therefore be less "efficient", in terms of information per subject, than if basketball players and comparison subjects had similar distributions of the risk factors being controlled. With the same total number of subjects and the same risk ratio, the study with more similar comparison groups will yield a narrower confidence interval (greater statistical precision) as well as a smaller p-value (greater statistical significance).

One way to obtain a better balance in risk factors between the basketball players and the comparison group is to match the comparison group to the basketball player group on the most important risk factors. For example, we could stratify the basketball players by height and GPA (grade point average) during college. A two-way stratification might have a total of 16 strata. We could then select comparison subjects so as to have the same distribution across these 16 strata. Choosing the comparison group in this way is called frequency or category matching. (This study might also be a logical place to use restriction, e.g., to include only males, aged 18-22 years, without any medical or physical impairments.)

The above method of frequency matching requires knowing the risk-factor distribution of the index group before enrollment of comparison subjects, so that the latter can be chosen to have the same distribution. Another method of accomplishing frequency matching is paired sampling. With paired sampling, a comparison subject is chosen sequentially or at random from among potential comparison subjects having the same covariable values as each index subject. For example, every time a former basketball player enrolls in the study, we find a comparison subject belonging to the same height-GPA stratum as the basketball player. Whenever we stop enrolling subjects, the two groups will have identical distributions across the 16 strata.

Similar to paired sampling is pair matching. In pair matching, we choose each comparison subject according to characteristics of an index subject, characteristics that not widely shared with other index subjects (i.e., strata are very small, possibly containing only one index subject each). For example, we might decide to use as comparison subjects the brothers of the index subjects. Or, we might decide that we wanted the joint height-GPA distribution to be so similar between player and comparison groups that we did not want to have to categorize the variables. In this case we would choose each comparison subject to have his height within a certain range (e.g., 2 centimeters) of the index subject's height and GPA within a certain small range of the index subject's GPA (pair matching in this way is called "caliper matching", though it has been criticized – see Rothman and Greenland).

What differentiates pair matching from paired sampling and other forms of frequency matching is the tightness of the link between index and comparison subjects. If there are multiple index-comparison subject pairs in each stratum, so that the pairs could be dissolved, shuffled, and reformed, with no effect as long as all subjects stayed in their strata, then the situation is one of frequency matching. If, in contrast, comparison subjects are for the most part not interchangeable with other comparison subjects, if each comparison subject is regarded as fully comparable only to the index subject with whom he is paired, then the situation is one of pair matching. (For a discussion of paired sampling versus pair matching, see MacMahon and Pugh, 1970, pp. 252-256. Also, although the present discussion has focused on pairs, all of these concepts apply to triplets, quadruplets, and "n-tuplets", as well as to variable numbers of comparison subjects for each index subject, e.g., the index subject's siblings.)

In case-control studies, as we saw earlier, the study architecture prevents us from ensuring that exposure groups are similar with respect to other risk factors even in the study population, and certainly not in the study base. Therefore, matching in a case-control studies does not prevent confounding. Matching can be beneficial, though, since if important potential confounders are similarly distributed in cases and controls, the comparison of these two groups can be more statistically efficient – with the same number of participants, the confidence interval for the odds ratio estimate will be narrower (i.e., the estimate will be more precise).

Unfortunately, the issue of whether or not it is beneficial to match controls to cases turns out not to have a simple answer, since in some cases matching can lead to reduced statistical efficiency. If the matching variable(s) are strongly associated with the exposure, then the exposure prevalence in matched controls will be more similar to that in cases than would occur for an unmatched control group, thereby diminishing the observed strength of association between exposure and disease. If the matching factors are not strong risk factors for the disease, then "overmatching" has occurred and a true association may be completely obscured.

The current advice for case-control studies is to match only on strong determinants of the outcome under study, especially if they are likely to be very differently distributed in cases and controls. Also, of course, do not match on a variable whose relationship to the outcome is of interest. Once you have matched cases to controls on a variable, its odds ratio will be one. Although matching in a follow-up study does not incur the problems that can arise in case-control studies, in any study design the use of matching can present practical and logistical difficulties, particularly if the pool of potential comparison subjects is small or if identifying or evaluating potential matches is costly.

Randomization

Randomization, the random assignment of participants to "exposed" or "treatment" and comparison groups, is available only in intervention trials. Randomization will ensure that, on the average, index and comparison groups will have similar proportions and distributions of all factors. Of course, in any particular study the groups may (and often will) differ in one respect or another (i.e., the randomization will not "work", though in a more precise sense, it does work – it just does not accomplish all that we would like it to). So often intervention and control groups will be constrained to be similar (through matching, also called "pre-stratification") or will be analyzed using stratified analysis.

An important consideration regarding randomization – and its decisive advantage over any of the other methods available – is that on the average randomization controls for the effects of variables that cannot be measured or are not even suspected of being risk factors. Unless a variable has been identified as relevant and can be measured, none of the other approaches described above (or below) can be used. With randomization, we have the assurance that at least on the average we have accommodated the influence of unknown and unsuspected risk factors.

Stratified analysis

Stratified analysis involves the breaking down of an aggregate into component parts so that we can observe each subcomponent individually. If smoking is a relevant factor for the disease under study, we simply say, "very well, we will look at the smokers and then we will look at the nonsmokers". Most of the examples of confounding and effect modification we have examined have been presented in terms of stratified analysis.

Stratified analysis is intuitively meaningful and widely used. It is particularly suited to the control of nominal variables (variables whose values have no ordered relation to one another, such as, geographical region [north, east, west]) and ordinal variables that have few categories (e.g., injury severity [minor, moderate, severe]). Stratified analysis gives a "picture" of what is going on in the data, is easily presented and explained, and requires no restrictive assumptions about a statistical model.

On the other hand, stratified analysis requires that continuous variables be categorized, which introduces a degree of arbitrariness and causes the loss of some information. It is not possible to control for more than a few variables at the same time because as the number of strata grows large, understanding and interpreting the results may present a major challenge, especially if the results vary from one stratum to another without any obvious pattern. Despite these drawbacks, stratified analysis is a mainstay of epidemiologic analysis approaches.

When there are multiple strata, it may be difficult to describe and to summarize the results, particularly since many strata will contain relatively few participants, so differences might readily be due to random variation. In such a case, various summary measures – generally different forms of weighted averages of the stratum-specific measures – are available. A summary measure is a single overall measure of association over all strata (or over a subgroup of the strata), controlling for the variables on which stratification has taken place. The standardized risk ratio (SRR) presented in the section on age standardization is one such summary measure. Others will be presented in the chapter "Data analysis and interpretation". Of course, as with any summary measure, if there are important differences across strata an overall average may not be meaningful.

Modeling

Given an unlimited number of participants, and an unlimited amount of time, patience, and capacity to interpret data, we could approach any multivariable analysis problem by means of stratification. But consider the dimensions of the challenge: if we have three variables, each dichotomous, there are eight possible unique strata; if we have six variables, each dichotomous, there are 64; if we have six dichotomous variables and three variables having three levels each, the number of strata soars to 1728! Imagine trying to interpret 1728 odds ratios, even assuming that we have enough participants for each one.

Since we often have more than a few variables we wish to accommodate, and variables (e.g., age, blood pressure, body weight) are often continuous so that we stand to lose information by categorizing them into any small number of levels, there is an obvious need for some more sophisticated approach that does not require us to examine every possible combination of factor levels in order to uncover the effects of each variable. There is such an approach – mathematical modeling – but its use involves a price, in terms of certain assumptions we make in the interests of simplifying the situation. Another price we pay is that the data themselves are hidden from view. In the words of Sir Richard Doll (interview with Beverly M. Calkins printed in the American College of Epidemiology Newsletter for Fall 1992):

"There have been many important steps along the way: larger scale studies, more powerful statistical techniques, and the development of computers that allow these techniques to be applied. I fear, however, that the ease of applying statistical packages is sometimes blinding people to what is really going on. You don't have a real close understanding of what the relationships are when you put environmental and all of the other components of the history together in a logistic regression that allows for fifteen different things. I am a great believer in simple stratification. You know what you are doing, and you really want to look at the intermediate steps and not have all of the data in the computer".

Limitations in the ability to control potential confounders

Typically, epidemiologists do not know all of the determinants of the health conditions they study. Other determinants may be known but cannot be measured, either in general or in the circumstances under study. Unknown and unmeasured potential confounders can be controlled only through randomization. This unique advantage of randomized designs is a primary reason for their particular strength.

Even for potential confounders that are controlled through restriction, matching, stratified analysis, or modeling, limitations or errors in the conceptualization, measurement, coding, and model specification will compromise the effectiveness of control. Such incomplete control results in "residual confounding" by the potential confounder. Residual confounding, like uncontrolled confounding, can lead to bias in any direction (positive or negative, away from the null or towards the null) in the adjusted measure of effect between the study factor and outcome. Even if measurement error in the potential confounder is nondifferential (i.e., independent of the study factor and outcome), the bias in the association of primary interest can be in any direction.

It is important to be aware of these limitations, but they are not grounds for discouragement. Notwithstanding these and other obstacles, epidemiology has provided and continues to provide valuable insights and evidence. The limitations derive primarily from the subject matter – health - related phenomena in free-living human populations – rather than from the discipline. Remaining aware of limitations, minimizing them where possible, and insightfully assessing their potential impact in interpreting data are the mark of the well-trained epidemiologist.

Confounding and effect modification

As noted in the chapter on Causal Inference, epidemiology's single variable focus, the one-factor-at-a-time approach that underlies the evolution of epidemiologic understanding, is the basis for the concepts of "confounding" and "effect modification". There are also some similarities in the way that they are investigated in data analysis. To make the distinction clear, we will contrast these two different implications of multicausality.

If we observe an association between a disease and some new factor - but fail to adequately account for possible effects of known causes of the disease - we may erroneously attribute the association we observe to the new factor when in fact we may be seeing the effects of known factors. "Confounding" refers to a situation in which an observed excess of disease can be mistakenly attributed to the exposure of interest when, in fact, some other factor – related to both the outcome and the exposure – is responsible for the observed excess. For example, the crude death rate in Florida is higher than in Alaska. If we attribute the higher death rate in Florida to the effect of citrus fruit industry, then we have fallen afoul of confounding. For the underlying "true" reason for the higher Florida death rates is the older age distribution of the Florida population.

When considering confounding, we are asking the question "Is the observed association between oral contraceptive use and myocardial infarction risk due to an effect of oral contraceptives or is the association actually due to the effects of other MI risk factors, such as cigarette smoking, elevated blood pressure, elevated blood cholesterol, and diabetes, that happen to be associated with oral contraceptive use?" To answer that question, we will attempt to ascertain that the groups being compared are the same with regard to these "potential confounders" and/or we will examine the OC-MI relationship within categories of the "potential confounders" in an attempt to "hold other factors constant".

"Effect modification" refers to variation in the relationship between exposure and outcome, variation that is due to the actions of some other factor (called an effect modifier). For example, the relationship between exogenous estrogens and endometrial cancer appears to be weaker in the presence of obesity. The relationship between oral contraceptives and myocardial infarction appears to be stronger in women who smoke cigarettes than in those who do not.

When considering effect modification, we are asking the question "Is the observed association between oral contraceptive use and MI risk importantly influenced by other MI risk factors, such as cigarette smoking, elevated blood pressure, elevated cholesterol, or even by factors which, by themselves, do not affect MI risk?" To answer that question, we will examine the OC-MI relationship within categories of these "potential modifiers". We will also seek biological and/or behavioral explanations for possible modifying influences.

With confounding, we are concerned with determining whether a relationship between our exposure and our outcome does or does not exist. With effect modification, we are concerned with defining the specifics of the association between the exposure and the outcome. That is, we are interested in identifying and describing the effects of factors that modify the exposure-outcome association. The question about confounding is central in establishing risk factors. The question about effect modification has important implications for defining disease etiology and for intervention. Confounding is a nuisance. Effect modification, though for statistical reasons it may be difficult to assess, is of considerable potential interest.

A mnemonic aid that may be helpful is the following. An evaluation of confounding is an investigation into "guilt" or "innocence". An evaluation of effect modification is an investigation into "conspiracy".

MAIN POINTS

- Confounding is a distortion or misattribution of effect to a particular study factor. It results from noncomparability of a comparison group.
- A confounder is a determinant of the outcome or its detection, or possibly a correlate of a determinant, that is unequally distributed between groups being compared.
- A determinant of the disease should appear as an independent risk factor, i.e., not one whose association with disease results from its association with the study factor.
- A potential confounder (i.e., a disease determinant) need not be an actual confounder – an actual confounder must be associated with the study factor.
- Confounding can be controlled in the study design and/or analysis.
- Control through the study design is accomplished through restriction, matching (prestratification), or randomization.
- Control in the analysis is accomplished through stratified analysis and/or mathematical modeling.
- Adequacy of control is compromised by errors in the conceptualization, measurement, coding, and model specification for potential confounders.
- Confounding deals with "guilt" or "innocence"; effect modification deals with "conspiracy".
- Discovery that an association arises from confounding does not make it less "real", but does change its interpretation.
- The crude association is real and for some purposes is the relevant measure.

Bibliography

Rothman and Greenland (see index); Rothman, *Modern epidemiology*, pp. 177-181 and 226-229.

W. Dana Flanders and Muin J. Khoury. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. *Epidemiology* 1990; 1:239-246.

Greenland, Sander; Hal Morgenstern, Charles Poole, James M. Robins. Re: "Confounding confounding". Letter and reply by D.A. Grayson. *Am J Epidemiol* 1989; 129:1086-1091

Savitz, David A.; Anna E. Baron. Estimating and correcting for confounder misclassification. *Am J Epidemiol* 1989; 129:1062-1071.

Mickey, Ruth M; Sander Greenland. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989; 129:125-37.

Stellman, Steven D. Confounding. *Preventive Medicine* 1987; 16:165-182 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Greenland, Sander and James M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 1986; 15:412-418. (advanced)

Schlesselman, J.J.: Assessing effects of confounding variables. *Am J Epidemiol* 108(1):3-8, 1979.

Schlesselman, J.J. *Case-control studies*. Pp. 58-63.

Kleinbaum, Kupper and Morgenstern. *Epidemiologic research: principles and quantitative methods*. Chapter 13, Confounding.

Boivin, Jean-Francois; Sholom Wacholder. Conditions for confounding of the risk ratio and of the odds ratio. *Am J Epidemiol* 1985; 121:152-158.

Greenland, Sander; James M. Robins. Confounding and misclassification. *Am J Epidemiol* 1985; 122:495-506.

Kupper, Larry L. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 1984; 120:643-8.

Greenland, Sander. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564-9.

Greenland, S. and Neutra R.: Control of confounding in the assessment of medical technology. *International J Epidemiology* 9:361-367, 1980.

Race/ethnicity

Amott T, Matthaei J. *Race, gender, and work: a multicultural economic history of women in the United States.* Boston, South End Press, 1991.

Crow JJ, Escott PD, Hatley FJ. *A history of African Americans in North Carolina.* Raleigh, Division of Archives and History, N.C. Department of Cultural Resources, 1992.

Franklin, John Hope. *From slavery to freedom: a history of African Americans.* 7th ed. NY, McGraw-Hill, 1994.

Freeman HP. The meaning of race in science – considerations for cancer research. *Cancer* 1998;82:219-225.

Gamble, Vanessa N. Under the shadow of Tuskegee: African Americans and health care. *Am J Public Health* 1997;87:1773-.

Greenland, Sander; Judea Pearl, James M. Robins. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37-48.

Hacker A. *Two nations: black and white, separate, hostile, unequal.* NY: Macmillan; 1992.

Kaufman, Jay S.; Sol Kaufman. Assessment of structured socioeconomic effects on health. *Epidemiology* 2001;12:157-167.

Krieger ND, Rowley DL, Herman A. Racism, sexism and social class: implications for studies of health, disease, and wellbeing. *Am J Prev Med* 1993;9(6[Suppl]):82-122.

Morbidity/mortality gap: is it race or racism? American College of Epidemiology Tenth Annual Scientific Meeting, *Ann Epidemiol* 1993;3:119-206.

Moss, Nancy. What are the underlying sources of racial differences in health? Editorial. *Ann Epidemiol* 1997;7:320.

Osborne NG, Feit MD. The use of race in medical research. *JAMA* 1992;267:275-279.

Stepan N. *The idea of race in science.* London, Macmillan, 1982.

Senior P, Bhopal RS. Ethnicity as a variable in epidemiologic research. *British Med J* 1994;309:327-330.

Varma J. Eugenics and immigration restriction: lessons for tomorrow. *JAMA* 1996;275:734.

Warnecke, Richard B., et al. Improving question wording in surveys of culturally diverse populations. *Ann Epidemiol* 1997;7:334-.

Williams, David R. Race and health: basic questions, emerging directions. *Ann Epidemiol* 1997;7:322-333.

Appendix

The following discussion is for the more advanced student (either now or when you are taking a more advanced methodology course) – others can skip this section.

Confounding: "Comparability" versus "collapsibility"

As presented earlier, the comparability definition labels as confounding a situation where the distribution of an outcome for the unexposed group differs from the (contrafactual) distribution for that outcome in the exposed group if it could be observed without the exposure. The collapsibility definition sees confounding as a situation where the crude measure of association differs from the value of that measure when extraneous variables are controlled (by stratification, adjustment, or mathematical modeling). The two definitions yield the same judgment in many situations, a major exception being those where the measure of association is an odds ratio which does not estimate a risk ratio (rare disease assumption not met) or a rate ratio (assumptions for estimating the IDR not met).

The reason the odds ratio is different from the rate and risk ratios in this respect is related to the fact that unlike proportions and rates, the odds for a group are not a simple average of individual members' odds (Greenland S. *Am J Epidemiol* 1987;125:761). Stratified analysis simply places the individual members of a group into a handful of strata. Since incidence for a group does equal the simple average of the risks (or "hazards") for the individual members, the overall incidence (in exposed, unexposed, or overall) will also equal the average of the stratum-specific risks or rates, now weighted by the stratum size (number exposed, number unexposed, or total) as a proportion of the total (i.e., the distribution of participants across the strata).

For risk or rate, therefore, the comparison (by means of a ratio or difference) of overall incidence in the exposed to overall incidence in the unexposed is a comparison of weighted averages. If the weights in the exposed and unexposed groups are the same, then the comparison is valid (i.e., no confounding). In this case, the overall incidence ratio (or difference) is a weighted average of the incidence ratios (or differences) across the strata, a condition for nonconfounding proposed by Boivin and Wacholder (*Am J Epidemiol* 1985;121:152-8) and implies collapsibility. Since the weights are the distributions of exposed and unexposed participants across the strata, equal weights mean identical distributions, which in turn means that exposure is unrelated to the risk factor used to create the strata.

If the distributions of exposed and unexposed participants across strata differ (i.e., the exposure is related to the stratification variable), then the overall incidence in exposed and in unexposed participants are averages based on different weights, so their ratio and difference will not be equal to a weighted average of the stratum-specific incidence ratios and differences. Comparability and collapsibility are therefore not present, and the comparison between overall incidences is confounded by the stratification factor. However, since the odds for the group is not a simple average of the odds for individual members, none of the above holds for the odds ratio unless it is sufficiently rare that it approximates the risk ratio or has been obtained from a design that causes the odds ratio to estimate the incidence density ratio.

Some of the relationships just presented can be readily demonstrated using simple algebra. Let a_i, b_i, c_i, d_i, n_{1i}, and n_{0i} in each stratum take on the values implied by the table below, and let their respective totals across all strata by a, b, c, d, n_1, and n_0 (i.e., a = all exposed cases, b = all unexposed cases, c = all exposed noncases, d = all unexposed noncases, n_1 = all exposed persons, n_0 = all unexposed persons).

	Exposure			
Disease	Yes	No	Total	
Yes	a_i	b_i	m_1	$(a_i + b_i)$
No	c_i	d_i	m_2	$(c_i + d_i)$
Total	n_{1i}	n_{0i}	n_i	
	$(a_i + c_i)$	$(b_i + d_i)$		

The incidence in exposed persons is a_i/n_{1i} within each stratum and a/n_1 when the strata are ignored (i.e., the total, or crude table). The (weighted) average incidence in the exposed across the strata is:

$$\Sigma \left(\frac{n_{1i}}{n} \times \frac{a_i}{n_{1i}} \right) = \Sigma \left(\frac{a_i}{n} \right) = \frac{a}{n}$$

where the summation goes over all strata. a/n is simply the crude incidence in the exposed. Similarly, the weighted average of the stratum-specific risk ratios can be expressed as the sum across all strata of:

$$\frac{w_i}{W} \times \frac{a_i / n_{1i}}{b_i / n_{0i}} = \frac{w_i}{W} \times \frac{a_i\, n_{0i}}{b_i\, n_{1i}}$$

where w_i are the weights for each stratum and W is the sum of the w_i. If we let $w_i = b_i n_{1i} / n_{0i}$, then we have the sum across strata of:

$$\frac{b_i\, n_{1i} / n_{0i}}{W} \times \frac{a_i\, n_{0i}}{b_i\, n_{1i}} = \Sigma \left(\frac{a_i}{W} \right) = \frac{a}{W}$$

Meanwhile, W is the sum across all strata of:

$$w_i \quad = \quad \frac{b_i\, n_{1i}}{n_{0i}}$$

If exposure is unrelated to the stratification variables, so that the distribution of exposed n_{1i}/n_1 is the same across strata as the distribution of the unexposed n_{0i}/n_0, then the ratio of exposed to unexposed in all strata must be the same as in the overall table, n_1/n_0. Therefore

$$w_i = \frac{b_i\, n_{1i}}{n_{0i}} = \frac{b_i\, n_1}{n_0}\ , \quad \text{whose sum is simply} \quad W = \frac{b\, n_1}{n_0}$$

$$\text{Thus, the sum of} \quad \frac{a_i}{w_i} \quad \text{is} \quad \frac{a}{bn_1\,/\,n_0}\ , \text{ which equals}$$

$$\frac{a\,/\,n_1}{b\,/\,n_0}\ , \text{ the overall risk ratio.}$$

So, when there is no confounding, the following three summary measures are all equal:

$$\text{Overall risk (or rate) ratio} \quad = \quad \frac{\text{Overall incidence in exposed}}{\text{Overall incidence in \underline{un}exposed}}$$

$$= \quad \frac{\text{Weighted average of incidence in exposed, across strata}}{\text{Weighted average of incidence in \underline{un}exposed, across strata}}$$

$$= \quad \text{Weighted average of stratum-specific risk (or rate) ratios}$$

With incidence odds and odds ratios, however, the above does not apply. The overall incidence odds are simply a/c. In contrast, the average of the stratum-specific odds, weighted by the number of exposed, is the sum over all strata of:

$$\frac{n_{1i}}{n_1} \times \frac{a_i}{c_i}$$

It is possible to construct an incidence odds ratio that is a weighted average of the stratum-specific incidence *odds ratios*, and therefore a summary incidence odds ratio. However, this summary incidence odds ratio will not be equal to a ratio of average stratum-specific incidence *odds* for exposed and average stratum-specific incidence *odds* for unexposed.