12

Analysis of Epidemiologic Studies: Evaluating the Role of Confounding

In Chapters 10 and 11, we discussed methods for evaluating the effects of two possible alternative explanations that must always be considered in assessing the presence of a valid statistical association in a given study—the role of chance and that of bias. The third alternative explanation, confounding, involves the possibility that the observed association is due, totally or in part, to the effects of differences between the study groups other than the exposure under study that could affect their risk of developing the outcome of interest. The concept of confounding is central to the interpretation of the findings of any epidemiologic study—most critically in observational studies but also for experimental investigations. Unlike bias, which is primarily introduced by the investigator or study participants, confounding is a function of the complex interrelationships between various exposures and disease.

THE NATURE OF CONFOUNDING

Intuitively, confounding can be thought of as a mixing of the effect of the exposure under study on the disease with that of a third factor. This third factor must be associated with the exposure and, independent of that exposure, be a risk factor for the disease. In such circumstances the observed relationship between the exposure and disease can be attributable, totally or in part, to the effect of the confounder. Confounding can lead to an overestimate or underestimate of the true association between exposure and disease and can even change the direction of the observed effect. For example, consider a study that showed a relationship between increased level of physical activity and decreased risk of myocardial infarction (MI). One additional variable that might affect the observed magnitude of this association is age. People who exercise heavily tend to be younger, as a group, than those who do not exercise. Moreover, independent of exercise, younger individuals have a lower risk of MI than older people. Thus, those who exercise could have a lower risk

of MI quite apart from any effect of this habit simply as a consequence of the greater proportion of younger individuals in this group. In this circumstance, age would confound the observed association between exercise and MI and result in an overestimate of any inverse relationship. Similarly, differences in the proportions of men and women could also potentially affect the magnitude of' the observed association between exercise and MI. A high level of exercise is likely to be more common in men, and, independent of exercise, men have a greater risk of MI than women. Thus, an inverse effect of exercise on risk of MI would be underestimated if differences in gender between exercisers and nonexercisers were not taken into account.

As can be seen from these examples, in the most general terms, for a variable to confound a relationship, it must be associated with both the exposure and the disease. If there is no association between the exposure and the potential confounder, or conversely, if the potential confounder has no relationship with risk of the disease, there can be no confounding by that factor. For example, those who exercise and those who do not are almost certain to differ with respect to their total daily consumption of fluids. Increased intake of fluids, however, has not in itself been shown to increase (or decrease) risk of MI. Thus, a difference in the level of this variable between those who are physically active and those who are not cannot be responsible for any decreased risk of MI observed among exercisers, and thus it is not a confounder of this association.

This general description of the characteristics of a confounder requires a number of refinements [16, 24]. First, while the potential confounding factor must, by definition, be predictive of the occurrence of disease, the association need not be causal. In fact, most frequently, confounding variables are only correlates of another causal factor. For example, age and sex are associated with virtually all diseases and are related to the presence or level of many exposures. Thus, they should always be considered as potential confounders of an association. These variables, however, may not be causally related to disease but rather act as surrogates for etiologic factors. The lower rates of coronary heart disease among women compared with men rnay not be due to gender, per se, but rather to correlates of sex, such as levels of endogenous hormones, that are more difficult to define and quantify.

Second, the potential confounding factor must be predictive of disease independently of its association with the exposure under study. In other words, the confounding factor cannot be related to risk of disease only through its association with the exposure. This means that there must be an association between the confounder and disease even among nonexposed individuals. In the previous example, if increased physical activity does, in fact, decrease risk of MI, then high levels of fluid consumption will also be associated with a decreased risk of MI, simply because fluid intake is associated with physical activity. However, fluid

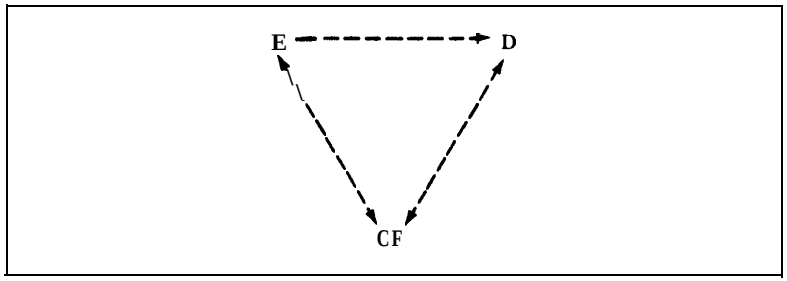

Fig. 12-1. Interrelationship between an exposure (E), confounding factor (CF), and disease (D).

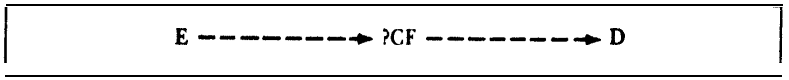

Fig. 12-2. Interrelationship between an exposure (E), disease (D), and a potential confounding factor (?CF) which is in the causal pathway and thus not a confounder.

consumption has not been shown to be associated with risk of MI among those who are physically inactive. Thus, this variable would not be a confounder of the association. This is clearly a different situation than for true potential confounders such as age, sex, and smoking, which not only are associated with physical activity, but, even among individuals who are not physically active, are risk factors for MI.

Finally, the potential confounder cannot merely be an intermediate link in the causal chain between the exposure and disease under study. This distinction is often not clear and requires knowledge or postulation of biologic mechanisms underlying the relationship of the exposure with disease. As shown in Figure 12-1, a confounder is a variable that is associated with the exposure and, independent of that exposure, is a risk factor for the disease. If, however, as shown in Figure 12-2, one mechanism of action of the exposure is to alter the level of the potential confounder, which in turn itself affects disease risk, then that factor is not a confounder but rather an intermediate step in the causal chain between the exposure and disease. For example, in evaluating the effect of moderate alcohol consumption on decreasing risk of MI, one variable that might at first glance be considered as a potential confounder is level of high-density lipoprotein (HDL) cholesterol. Studies have shown that alcohol raises HDL levels, and high levels of HDL are associated with reduced risk of MI independent of alcohol consumption [IO]. However, it has also been postulated that one mechanism for the effect of moderate alcohol consumption on risk of MI may be that it is mediated totally or

in part by changes in HDL. If such a mechanism is known or even assumed, then HDL would not be a confounder and should not be controlled for in the analysis. On the other hand, if it is of interest to assess the extent to which alcohol has an effect on MI by mechanisms other than by altering HDL levels, the analysis would need to control for HDL. Thus, level of HDL could be considered in different ways in separate analyses, depending on the specific research question as well as the known or postulated biologic mechanism.

As the foregoing discussion illustrates, the choice of the specific factors that should be considered as potential confounders of a particular association is often difficult. A practical means to determine whether a given factor is in fact an actual confounder in a study is to analyze the data, obtain a crude overall estimate of the association, control for the effect of the variable, and observe whether the estimate of the association between the exposure and disease is altered. A potential confounder would be an actual confounder if adjustment for the variable results in a change in the estimate of the association between the exposure and disease. It is important to note that in fact, the effect of any single confounder must be considered in the context of the effects of all other confounders in that study, since the ultimate concern is with the aggregate amount of confounding. However, long before that information is available, it is necessary in the design stage of the study to select variables that will be considered as potential confounders to ensure that adequate data are collected, since it is impossible to control later for the effects of a variable on which information was not obtained. The problem is that it is not logistically feasible to collect information on every variable that could possibly be related to either the exposure or the disease. Making a judicious choice about potential confounders depends on knowledge of the disease, previous evaluations of the same or related questions, and most importantly, the best judgment of the investigator at the time the study is initiated. Identification of a potential confounding factor is not a matter of determining the statistical significance of an association with the exposure or disease. As discussed in Chapter IO, statistical significance is a function of both the magnitude of an association and the sample size of the study. In a small sample, a confounding variable can certainly affect the magnitude of an observed association even if it does not achieve statistical significance in its relationship with either the exposure or disease. Conversely, in a very large study, a factor may be statistically significantly associated with exposure or disease even though it has only a small effect on the magnitude of the association. Thus, statistical significance is not the criterion on which to base the decision that a particular factor is or is not a potential confounder.

From a practical standpoint, since a confounder must be associated with both the exposure and the disease and it is often difficult to know what factors are correlated with the exposure, investigators may attempt to ensure that at the very least data are obtained on all available risk factors for the disease under study. For example, in a study of exercise and MI, it is not clear which demographic, medical, and life-style variables are associated with varying levels of physical activity. Since a number of specific factors have been shown consistently to affect risk of MI, at a minimum information on each of these could be collected. Moreover, the data collected need to be sufficiently detailed to permit adequate control of the effects of each variable. For example, to control the potential confounding effects of cigarette smoking on the association between exercise and MI, it would not be sufficient to define smoking status as simply ever versus never smoked, because risk of MI is associated with current rather than past habits. Since the group of ever smokers would include those who smoked currently, some control of confounding by this variable would be achieved; however, residual confounding due to this less than optimal categorization of smoking status would likely still affect the estimate of the magnitude of the association between exercise and MI. Since uncontrolled confounding is a major threat to the validity of results, it is imperative that the design features of the study permit the collection of adequate data to address this issue.

In assessing the effect of a potential confounding factor, it is important not merely to evaluate its presence or absence, but also to identify the direction and quantify the magnitude of its effect on the estimate of the association between the exposure and disease. The magnitude of the effect of the confounding present will depend on the magnitudes of the specific associations between the confounder and the exposure, as well as between the confounder and the disease. The direction of the effect of the confounding factor on the estimate of the observed association will depend on the nature of the interrelationships among the exposure, confounding factor, and disease. Positive confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is more extreme—either more positive or more negative-than the true association. For example, in a study of coffee consumption and MI, cigarette smoking would be a potential positive confounder since those who drink coffee are more likely to smoke and those who smoke are at an increased risk of MI. It therefore could appear that coffee drinkers have a higher risk of MI than they actually do, simply because they are also cigarette smokers, whose risk of MI is independently elevated. Analogously, smoking could also be a potential positive confounder of the protective association between physical activity and MI. Since those who exercise heavily tend to smoke less than those who do not and those who smoke less have a lower risk of MI than heavy smokers, it could appear that high levels of physical activity are even more protective than they

actually are simply because of the lower proportion of smokers among those who are physically active. In contrast, negative confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is an undercstirnatc of the true association. For physical activity and MI, for example, gender would be a potential negative confounder since being male is associated with increased level of physical activity but also with increased risk of MI. Consequently, a group of exercisers would contain a greater proportion of males than a comparison group of nonexercisers. This would result in a higher risk of MI among those who are physically active compared with those who are not due to the effect of the confounder. The observed magnitude of the association between physical activity and MI would therefore be closer to the null value than is actually true, thereby tending to underestimate any protective effect of physical activity on MI.

Note that the terms *positive* and *negative* in this context refer to the effects of the confounding factor on the direction of the observed risk estimate relative to the true parameter, regardless of the direction of the true effect of exposure on disease. Failure to control for negative confounding results in an observed estimate of effect that is diluted towards the null, meaning that any true increased or true decreased risk will be underestimated and appear as less of an association than is actually the case. Similarly, failure to control for positive confounding results in a more extreme estimate of effect observed than is actually the case. This can be in the direction of either an apparently stronger increased risk or an apparently more protective effect, depending on the direction of the true association.

An understanding of the direction in which a potential confounder is likely to affect the association between an exposure and disease can be very inforrnative in situations where the factor was not or could not be controlled. For example, in one of the earliest analytic studies evaluating the relationship of smoking with lung cancer, Doll and Hill [6] found a highly significant difference in the percentage of smokers between cases of lung cancer and controls of the same age and sex with diseases other than cancer. One concern in interpreting those results was that the cases and controls differed with respect to place of residence. Specifically, a higher proportion of the lung cancer patients resided in rural areas outside of London at the time of their admission to the hospital. Since there were also urban/rural differences with respect to smoking patterns, place of residence fulfilled the requirements for being a potential confounder. The real question, however, is whether the uncontrolled effect of differences in place of residence between cases and controls could have accounted for the observed increased risk of lung cancer associated with smoking. The relationship of place of residence with smoking was such that rural areas had a lower smoking rate than urban areas. Rural place of residence would therefore be a negative confounder in these data, since cases had a higher proportion of those from rural areas, those from the rural areas-were less likely to smoke, and those less likely to smoke were less likely to develop lung cancer. 'Thus, in these data, urban/rural differences could only act to underestimate any true association between smoking and lung cancer and could not have been responsible for the observed increased risk.

There are a number of methods that can be employed, singly or in combination, to control for the effects of confounding in analytic epidemiologic studies. These include randomization and restriction, which are features adopted in the design phase of a study; matching, which involves both study design and analysis; and two specific analytic techniques, stratification and multivariate analysis. The basic strategy underlying all of these methods derives directly from the definition of confounding. A factor can confound an association only if it differs between the study groups. As mentioned in the example of exercise and risk of MI, gender would be a confounder only if there were different proportions of men and women among those who exercise and those who do not and if, in addition, gender were an independent risk factor for the disease. If gender did not vary between exercisers and nonexercisers, it could not be a confounder of the association between exercise and MI. Thus, if through the way we design or analyze the study, we make comparisons only among individuals with the same level of the confounding variable, then the confounding effect of that factor will be controlled. If we examine the association between exercise and MI among only men, or among only women, or among both men and women but separately, then gender will not vary within each of these groups of subjects, and the resultant estimate of the association will not be confounded by this variable.

METHODS TO CONTROL CONFOUNDING IN THE DESIGN

There are three methods that can be used to control confounding in the design of analytic epidemiologic studies: randomization, restriction, and matching. Randomization is applicable only to intervention studies, while restriction and matching can be considered for all analytic study designs.

Randomization

Randomization, as discussed in detail in Chapter 8, is for many reasons the procedure of choice in an intervention study for the allocation of study subjects to the various exposure categories. The unique strength of randomization relates to its ability to control confounding. With a

sufficient sample size, randomization will virtually ensure that all potential confounding factors-those known to the investigator and, even more importantly, those currently unknown or even as yet unsuspected-are evenly distributed among the treatment groups. This ability of randomization to control for unknown confounders cannot be achieved by any other approach in the design or analysis of' an epidemiologic study. If there are imbalances in known or suspected risk factors due to small sample size, or to the play of chance even in a large sample, these can be controlled in the analysis using the techniques that will be discussed below. Of course, imbalances with respect to unknown risk factors can never be controlled in the analysis. Thus, when using randomization to control confounding, it is important that the sample size of the study be sufficiently large.

Restriction

As stated earlier, confounding cannot occur if the potential confounding factors do not vary across either the exposure or the disease categories. One way to achieve this is to restrict the admissibility criteria for subjects and limit entrance into the study to individuals who fall within a specified category or categories of the confounder. For example, if sex and race are potential confounding factors, the study could include only nonwhite men or only white women. Similarly, control of age could be achieved by restricting admissibility to those within a narrow range that corresponds to a relatively homogeneous rate of disease incidence. To the extent that variation in the confounding factor can be either eliminated (as in the case of race and sex) or substantially reduced (as in the case of age) by restricting the study population, the confounding effect of that variable is correspondingly eliminated or minimized.

Restriction is a straightforward, convenient, and inexpensive means to control confounding. If the range of permissible values of a potential confounder is sufficiently narrow, restriction offers virtually complete control. There are, however, a number of limitations to the use of restriction that should be considered. First, the use of restriction may substantially reduce the number of subjects eligible to participate in the study, which may present difficulties in achieving the sample size necessary for adequate statistical power in a reasonable period of time. Second, the use of restriction involves the potential for residual confounding if the criteria are not sufficiently narrow. For example, in a study of physical activity and Ml, one important potential confounding factor that might be dealt with by restriction would be age. If, however, the study population were restricted to the category of those aged 40 to 65, it is certain that there would be residual confounding, because the rates of both MI and physical activity vary widely within that broad age range. Similarly, restricting the study population to individuals who had ever smoked cigarettes would not be sufficient to control adequately for the effects of smoking because risk of MI is most strongly associated with current, not past smoking.

Finally, perhaps the most serious potential disadvantage is that while restriction can deal effectively with the effects of a confounding variable, it does not permit evaluation of the association between exposure and disease for varying levels of the factor. For example, in assessing the association between exercise and risk of MI, restricting the study population to either men or women would certainly eliminate any effect of sex as a confounding factor. On the other hand, it might also be of interest to know whether the existence or magnitude of the association between exercise and MI differs between men and women. Clearly. this question could not be evaluated directly in a study that restricted the population to only one sex. It must be kept in mind that restriction may limit generalizability but in no way affects the validity of any observed association between the groups that were included in the study. Indeed, restriction enhances validity by providing an estimate that is unconfounded by the restriction factors.

Matching

Unlike randomization and restriction, which are used to control confounding in the design stage of a study, matching is a strategy that must include elements of both design and analysis. With restriction, the control of confounding is achieved by selecting into the study only individuals with certain homogeneous levels of the potential confounders. With matching, all levels of these factors are allowable for inclusion in the study, but the particular subjects are selected in such a way that the potential confounders are distributed in an identical manner among each of the study groups. For example, in a case-control study of exercise and risk of MI in which age, sex, and smoking are potential confounders, for each case of MI a control would be selected of the same age, sex, and level of smoking. Specifically, a 65-year-old female MI patient who was currently a heavy cigarette smoker would be matched to a woman of the same age and smoking status who had never suffered an MI. In this way, matching forces the distribution of these potential confounders to be identical in both study groups. When matching in the design is combined with the appropriate analysis, as discussed below, control of confounding by the matching factors is achieved.

Matching as a technique for the control of confounding has great intuitive appeal and has been widely used over the years; however, it also has a number of logistic and scientific disadvantages. These disadvantages, coupled with the more recent development of alternative techniques for the control of confounding in the analysis, have lessened the desirability of and necessity for matching except in certain clearly de-

fined circumstances. The first disadvantage of matching is that it can be difficult, expensive, and time-consuming to find a comparison subject with the right set of characteristics with respect to every matching variable for each individual enrolled in a study. Thus, although in principle matching can be used in any analytic study design, it is rarely used in large-scale cohort studies. In such studies it is usually more cost-effective simply to admit a large pool of comparison subjects and use other methods for controlling confounding, such as stratification or multivariate analysis. One exception to this rule is when there are a very large number of comparison subjects available and their data are easily and cheaply accessible, such as by computer tape. In this circumstance, the costs with respect to time and money in identifying matched controls will be negligible.

Thus, matching is primarily utilized in case-control studies, which in general tend to be smaller in sample size. Even here, the cost of obtaining information on potential confounders and selecting matched controls must be taken into account. It can often be expensive and time-consuming to assemble a suitable study population if matching is employed. For example, even if there are only three factors that are to be matched on-age (in five categories), sex (in two categories), and race (in three categories)-there are already 30 (5 x 2 x 3) possible combinations that may have to be considered in finding an appropriate control. A number of potential controls may have to be excluded before finding one with the particular set of characteristics of the case. When there are sufficient numbers of cases available for study, a 1 : 1 match (one control per case) provides the most statistically efficient design. When the number of cases is limited or fixed, however, the statistical power of the study to detect an association if one truly exists can be increased by selecting more than one control per case (referred to as R : 1 matching). This is an especially attractive alternative when the cost of obtaining information from cases and controls is similar. In general, however, as the ratio of controls to cases increases beyond 4 : 1, the additional gain in statistical power for the crude comparison may be small compared with the costs in terms of both time and money [18].

In addition to these disadvantages in terms of time, money, and loss of potential study subjects, as was the case with restriction, another limitation of matching is the inability to evaluate the effect of a factor that has been matched on the risk of the outcome. Thus, the effect of the confounding factor itself on risk of disease, such as the effect of smoking on risk of MI in a case-control evaluation of exercise and MI where smoking levels have been matched, cannot be explored because the distribution of that factor has been forced to be identical between cases and controls. Moreover, the use of matching does not control potential confounding by factors other than those for which matching was done, except indirectly for factors highly correlated with the matching variables. Matching may, in fact, result in greater difficulty in controlling for additional confounders. Stratified analysis cannot easily be used with matched data to control for additional confounders that were not included as matching factors. Although they can be controlled using specialized multivariate techniques, including matched-pair conditional logistic regression [2 1,221, the effective sample size is reduced because the analyses are based on only discordant pairs, as will be discussed below. This can add both inefficiency and complexity that would not have been necessary if matching had not been utilized and the effects of all potential confounders had-been controlled in the analysis.

Despite these substantial scientific limitations and logistic difficulties, there are some circumstances in which matching is desirable and even necessary. First, for certain variables. if matching were not employed in the design phase of the study, there would not be a sufficient number of individuals in the study groups who were alike with respect to these confounding factors to allow for any type of control in the analysis. In other words, matching is necessary for any factors for which there would otherwise be insufficient overlap between the study groups. Complex nominal variables such as neighborhood or sibship, which represent a wide and undefinable range of environmental and genetic factors, are especially difficult, if not impossible, to quantify and thus control by other means. By matching each case to a sibling- control, for example, the investigator attempts to control for a number of general characteristics highly correlated with membership in a particular family, including genetic factors, early environmental exposures, dietary habits, socioeconomic status, and utilization of health care resources. In the same way, individuals from the same neighborhood are likely to have experienced similar environmental exposures and be more alike with respect to individual life-style variables as well as correlates of social class and ethnic group. If controls were selected at random from the general population and the association between the exposure and disease among those living in the same neighborhood were examined afterward, it is very likely that only one or two individuals would come from each neighborhood, making an analysis impossible. A control from that neighborhood would therefore need to be specially selected or matched to each case to ensure that the necessary comparable information is available. This would be true of any such variable with a number of possible values on a nominal scale.

A second circumstance in which matching may be useful is when the case series is very small. In this situation, baseline characteristics are very likely to differ between the study groups due to chance variability, and yet the sample size will simply not be sufficient to form adequate subgroups of the confounders to control such variables in the analysis.

Matching a number of controls to each case with respect to the potential confounders will ensure an adequate number of cases and controls for each of the subgroups so that it will be possible to evaluate efficiently the association between exposure and disease.

Taking these disadvantages and advantages into account, we believe that matching should not be used routinely, but only after careful consideration of its appropriateness for a particular study. The wide availability of alternative techniques for the control of confounding in the analysis has markedly decreased the necessity and desirability of matching except in certain clearly defined circumstances. In most situations, it seems far preferable to choose an adequate sample of the comparison group and control for confounding through stratification and/or multivariate analysis.

In considering the analysis of data from case-control studies that have utilized matching, it is important to understand the role of matching in the design on the control of confounding. By matching cases and controls for a number of potential confounders, the groups have been selected to be more alike with respect to these factors than would have occurred had two independent series of cases and controls been chosen. If the matching factors are true confounders, the matching will result in a greater similarity between cases and controls with respect to their exposure histories than would otherwise be the case. If this similarity produced by the matching is not taken into account in the analysis by utilizing statistical techniques that make explicit provision for the matched nature of the data, an underestimate of the true association between exposure and disease will result. Thus, matching in the design of a case-control study does not by itself control confounding. Rather, control of confounding results from matching in the design coupled with stratification in the analysis. The true utility of matching in the control of confounding relates primarily to considerations of analytic efficiency or the ability to test a hypothesis with adequate statistical power and thereby produce a precise estimate of effect. By ensuring an adequate number of cases and controls for each level of a given confounding factor, matching permits effective control in the analysis through the use of stratification or multivariate analytic techniques.

Figure 12-3 illustrates the presentation of data for the analysis of a matched-pair case-control study. Unlike the two-by-two table for unmatched data, in which each cell represents the number of individuals with a certain exposure and disease status, the cells for a matched study denote the number of pairs that fall into each category. In this table, the rows indicate the exposure status of the case, and the columns indicate the corresponding status of the control member of the matched pair. Thus, cell *a* indicates the number of pairs in which both the case and control are exposed, *b* the pairs in which the case is exposed but the control nonexposed, c those where the control is exposed but not the

		Control: Exposed	Control: Nonexposed	
Case	Exposed	*a*	*b*	*a* + *b*
	Nonexposed	*c*	*d*	*c* + *d*
		a + *c*	*b* + *d*	*T*

Fig. 12-3. Presentation of Data From a Matched Pair Case-Control Study in a 2 x 2 Table

case, and *d* the pairs in which neither the case nor the control is exposed. The margins indicate the total number of' pairs in which the case was exposed $(a + b)$ or nonexposed $(c + d)$, the total pairs in which the control was exposed $(a + c)$ or nonexposed $(b + d)$, and the total number of case-control pairs in the study $(a + b + c + d = T)$.

This presentation of matched-pair data can be illustrated using data from a study of postmenopausal estrogens and endometrial cancer [28]. As shown in Table 12- I, 3 17 cases of endometrial cancer were identified and matched to an equal number of women with other gynecologic neoplasms on the basis of age at diagnosis (within 4 years) and year of diagnosis (within 2 years). Exposure was defined as at least 6 months of estrogen use recorded before the diagnosis of' the cancer. There were 39 pairs in which both the case and her matching control used estrogens, as well as 150 pairs in which neither the case nor her matching control used estrogens. Since for these concordant pairs there was no difference between cases and controls with respect to their exposure, they provide no information on the magnitude of the association between estrogen use and risk of endometrial cancer. All such information is provided by the discordant pairs, that is, those in which one member was exposed and the other was not. In this study, there were 128 pairs discordant for estrogen use: 113 in which the case was exposed but not the control, and 15 in which the control reported estrogen use while the case did not.

In a matched-pair analysis, the estimate of the magnitude of the association between the exposure and disease is based entirely on the ratio of the discordant pairs. Specifically, the relative risk of disease associated with exposure is calculated as the ratio of the number of pairs in which the case is exposed and the control nonexposed to the number of pairs

Table 12-l. Data from a matched-pair case-control study of exogenous estrogens and endometrial carcinoma

	Control		
	Exposed	Nonexposed	Total
Case			
Exposed	39	113	152
Nonexposed	15	150	165
Total	54	263	317

$$RR = \frac{b}{c} = \frac{113}{15} = 7.5$$

$$\chi^2_{(1)} = \frac{(b-c)^2}{(b+c)}$$

$$= \frac{(113-15)^2}{(113+15)}$$

$$= 75.03$$

$$95\%\ CI = RR^{(1 \pm 1.96/\chi)}$$

$$= 7.5^{(1 \pm 1.96/8.66)}$$

$$= (4.72, 11.92)$$

Source: Data from D. C. Smith et al., Association of exogenous estrogen and endometrial carcinoma. N. *Engl. J. Med.* 293:1164, 1975.

in which the control is exposed but not the case. This calculation can be represented as follows:

$$RR = \frac{b}{-}$$

Cells a and *d*, which represent the concordant pairs in which case and control are similar with respect to the exposure of interest, provide no information for this calculation. The estimate of the relative risk of endometrial cancer for those who used postmenopausal estrogens is therefore as follows:

$$RR = \frac{b}{-}$$

$$= \frac{113}{15}$$

$$= 7.5$$

This indicates that compared with women who did not use estrogens, those who used these agents were seven and a half times more likely to develop endometrial cancer. This estimate of effect can then be tested for statistical significance using the McNemar test or chi-square test for matched-pair data [7]:

$$\chi^2_{(1)} = \frac{(b-c)^2}{(b+c)}$$

$$= \frac{(113-15)^2}{(113+15)}$$

$$= 75.03,\ P = 4.8 \times 10^{-18}$$

Corresponding confidence intervals can also be calculated from the standard formulas presented in Chapter 10. For example, using a test-based approach, the matched-pair estimates of the relative risk and chi-square can be utilized. Thus, for the example above:

$$95\%\ CI = RR^{(1 \pm 1.96/\sqrt{\chi^2})}$$

$$= 7.5^{(1 \pm 1.96/8.66)}$$

$$= 7.5^{(1 \pm 0.23)}$$

$$= (4.72, 11.02)$$

In this example, the estimate of the relative risk calculated from the matched-pair analysis differs appreciably from that which would have been calculated from the same data with the matched pairs not retained in the analysis. As shown in Table 12-2, the information on the cases and controls can be considered as if derived from totally independent samples by inserting the numbers from the margins of the matched table, which represent the total numbers of subjects in each exposure and disease category, into the appropriate cells of the unmatched table. Thus, there were a total of 634 (317 times 2) individuals in the original study: 3 17 cases and 3 17 controls. Of the 3 17 cases, 152 used estrogens (ignoring the status of the matched control) and 165 were nonusers. Similarly, of the 317 controls, a total of 54 used estrogens and 263 did not. The unmatched relative risk would then be calculated:

$$RR = \frac{ad}{bc}$$

$$= \frac{(152)\ (263)}{(54)\ (165)}$$

$$= 4.5$$

Table 12-2. **Unmatched analysis of data from a matched-pair case-control study of exogenous estrogens and endometrial cancer**

	Cases	Controls	Total
Case			
Exposed	152	54	206
Nonexposed	165	263	428
Total	317	317	634

$$RR = \frac{ad}{bc} = \frac{(152)(263)}{(54)(165)} = 4.5$$

Source: Data from D. C. Smith et al., Association of exogenous estrogen and endometrial carcinoma. N. *Engl. J. Med.* 293: *1164,* 1975.

The unmatched estimate of effect is different from that of the matched, **reflecting the fact that the matching variables were in fact confounders of the association between exogenous estrogens and endometrial cancer. Thus, the matching must be taken into account in the analysis to provide a valid estimate of the association between estrogen use and risk of this disease.**

On the other hand, if the relative risk estimates from the matched and unmatched data had been similar, then the matching did not introduce a similarity between the cases and controls with respect to exposure history, and the analyses can be reported using unmatched data. Tables 12-3 and 12-4 present findings from the Boston Area Health Study [3], a **case-control study of risk factors for** Ml. **Cases of first MI from six Bos**ton area hospitals were matched to controls of the same age (within 5 years), sex, and neighborhood of residence. One hypothesis of interest was the relationship of level of physical activity, represented by an index that measures total kilocalories of energy expended from stairs climbed, blocks walked, and recreational/leisure-time activities, with risk of MI. High-level exercise was defined as a physical activity index of 2500 kcal or more per day.

Table 12-3 presents the results of the matched-pair analysis. The relative risk estimate calculated from these data is 0.60, indicating that compared with individuals who do not exercise, those who have a high level of physical activity had a statistically significant 40-percent decreased risk of Ml. When the matching was ignored (Table 12-4), the relative risk estimate calculated was virtually identical, indicating that there was no material confounding of the association between physical activity and MI by the matching factors. Consequently, the matching need not be **retained in the analysis. As a result, the analyses did not have to be lim**ited to the cases and controls for whom a suitable match was available. The entire sample, in this instance a total of 366 cases and 423 controls,

Table 12-3. **Data from a case-control study of physical activity and myocardial infarction: matched-pair analysis**

	Control		
	Exposed *	Nonexposed	Total
Case			
Exposed*	115	59	174
Nonexposed	99	67	166
Total	214	126	940

$$RR = \frac{b}{c} = \frac{59}{99} = 0.60$$

$$\chi^2_{(1)} = \frac{(b - c)^2}{b + c} = \frac{(59 - 99)^2}{59 + 99} = 10.13, \; P = 0.0015$$

$$95\%\ CI = RR^{(1 \pm 1.96/\chi)} = 0.60^{(1 \pm 1.96/3.18)} = (0.44, 0.82)$$

*Exposed = physical activity index ≥ 2500/kcal/day.

Table 12-4. **Data from a case-control study of physical activity and myocardial infarction: unmatched analysis**

	Cases	Controls	Total
Exposure*			
Yes	174	214	388
No	166	126	292
Total	340	340	680

$$RR = \frac{ad}{bc} = \frac{(174)(126)}{(214)(166)} = 0.62$$

$$\chi^2_{(1)} = \frac{\left[a - \frac{(a + b)(a + c)}{T}\right]^2}{\frac{(a + b)(c + d)(a + c)(b + d)}{T^2(T - 1)}} = \frac{\left[174 - \frac{(388)(340)}{680}\right]^2}{\frac{(388)(292)(340)(340)}{680^2(679)}} = 9.59, \; P = 0.002$$

$$95\%\ CI = RR^{(1 \pm 1.96/\sqrt{\chi^2})} = 0.62^{(1 \pm 1.96/3.1)} = (0.46, 0.84)$$

*Exposed = physical activity index ≥ 2500/kcal/day.

could be included in subsequent analyses, potentially increasing the statistical power of the study.

The magnitude of the underestimate of effect that may be introduced by analyzing matched data as if they were unmatched relates to the degree of confounding by matching factors in the study. When there is little or no confounding by the matching factors, as in the above example; when the data have been frequency matched (in which the matching is done in groups so that the same proportions of cases and controls are in each stratum of the confounder); or when matching has been used primarily for convenience, as is often the case with age and sex, the matched and unmatched relative risk estimates will be virtually identical, and an unmatched analysis can be done. The advantage of doing this is that it will more easily permit control in the analysis of the matching factors as well as other confounding variables that were not considered in the matching process by using stratification or multivariate analysis techniques as described below. Unnecessary matching may, in fact, result in a loss of statistical efficiency relative to having selected an unmatched comparison series, especially in a case-control study [12, 17]. The magnitude of the loss in efficiency will depend on the interrelationships among the exposure, the disease, and the confounder. Matching on variables that are either weak risk factors for the disease or not risk factors at all will, in fact, result in a substantial reduction of information in the analysis.

The formulas provided in this section pertain to individual 1 : 1 matching. Analogous formulas for parameter estimation and testing of matched data with more than one control per case (R : 1 matching) with a variable control to case ratio, and with multiple exposure levels are also available [19, 24].

METHODS TO CONTROL CONFOUNDING IN THE ANALYSIS

Stratified Analysis

Stratification is a technique to control confounding in the analysis of a study that involves the evaluation of the association within homogeneous categories or strata of the confounding variable. If, for example, sex were a potential confounder, an estimate of the association between the exposure and disease would be calculated for men and for women separately. Each of these stratum-specific estimates is, by definition, unconfounded by sex, since there is no variability of the confounding variable within the stratum. Similarly, if race (defined as black, white, and other) were also a potential confounder, to control for both race and sex simultaneously, stratum-specific estimates would be calculated separately

Table 12-5. Data from a case-control study of physical activity and risk of MI, stratified by history of cigarette smoking

Smoking history	Physical activity index	Cases	Controls	Total	
Never smoker	2500 + kcals	41	84	125	
	< 2500 kcals	46	52	98	RR = 0.55
	Total	87	134	223	
Exsmoker 10 + years	2500 + kcals	41	80	121	
	< 2500 kcals	30	39	69	RR = 0.67
	Total	71	119	190	
Exsmoker < 10 years	2500 + kcals	22	34	56	
	< 2500 kcals	21	26	47	RR = 0.80
	Total	43	60	103	
Current smoker	2500 + kcals	86	68	154	
	< 2500 kcals	79	40	119	RR = 0.64
	Total	165	108	273	
Total	2500 + kcals	190	266	456	
	< 2500 kcals	176	157	333	RR = 0.64
	Total	366	423	789	

for six categories: black men, black women, white men, white women, and men and women of "other" race. Each of these estimates would similarly be unconfounded by sex and race. It is possible simply to report the unconfounded relative risk estimate for each stratum and calculate a confidence interval around each estimate. It is also useful, however, to calculate a single overall estimate of the association between exposure and disease, once the effect of the confounding factor (or factors) has been taken into account. A number of statistical methods are available to combine the results of the unconfounded stratum-specific values into a single overall unconfounded estimate, all of which calculate a weighted average of the stratum-specific estimates of effect. The choice of the particular weights depends on the characteristics of the data.

The use of stratified analysis to control confounding can be illustrated by the earlier example concerning the association between level of physical activity and risk of MI. Cigarette smoking might potentially confound the association, since it is an independent risk factor for MI, and may be related to level of physical activity. To control for this variable, the overall association between physical activity level and MI is first considered separately for each stratum of cigarette smoking. As seen in Table 12-5, increased level of physical activity is **protective for each** of the

four levels of smoking, with an estimated relative risk of MI associated with physical activity of 0.55 for those who never smoked, and 0.67, 0.80, and 0.64 for exsmokers of 10 + years, exsmokers of less than 10 years, and current smokers, respectively. Each of these stratum-specific relative risks is an estimate of the association between physical activity and risk of' MI that is unconfounded within that defined range of cigarette smoking.

A single summary estimate of the association between physical activity and risk of MI that is unconfounded by smoking can then be derived from the stratified data by calculating a weighted average of the stratum-specific estimates. When, as in the example above, the stratum-specifi relative risk estimates appear to be similar or uniform over the range of the confounding variable, they can each be considered as providing a separate estimate of the same value of the magnitude of the overall association, with the individual estimates varying merely because of sampling variability or random error. The similarity of the estimates can be judged either by "eyeballing" the data or by performing an appropriate test of statistical significance. In this circumstance, to calculate a weighted average, the most precise estimate of the overall effect will derive from giving the greatest weight to the stratum-specific estimates with the largest sample sizes and thus the smallest variability. Specifically, this can be accomplished by assigning weights to the stratum-specific values that are inversely proportional to the variance of each estimate. This method of calculating the most precise overall estimate of effect, assuming uniformity of the stratum-specific estimates, is often referred to as pooling.

A simple method for calculating a pooled summary relative risk estimate from a series of two-by-two tables was proposed by Mantel and Haenszel [15]. Using standard notation as reviewed in Table 12-6, the formula for the pooled estimate of the relative risk for a case-control study can be expressed as follows:

$$RR_{MH} = \frac{\sum \frac{ad}{T}}{\sum \frac{bc}{T}}$$

where the quantities in the numerator and denominator are summed separately over each of the individual strata. Analogous formulas for the pooled estimators of the relative risk for cohort studies with count and person-year denominators are provided in Table 12-7 [25]. Analogous formulas for the pooled estimate of the attributable risk are considered elsewhere [9].

Calculating a pooled estimate of the association between physical ac-

Table 12-6. Notation of **a two-by-two table**

CASE-CONTROL OR COHORT STUDY WITH COUNT DENOMINATORS

	Cases	Controls	Total
Exposed	a	b	a + b
Nonexposed	c	d	c + d
Total	a + c	b + d	a + b + c + d = T

Case-control: $RR = OR = \frac{ad}{bc}$

Cohort: $RR = \frac{I_1}{I_0} = \frac{a/(a+b)}{c/(c+d)}$

COHORT STUDY WITH PERSON-YEARS DENOMINATORS

	Cases	Controls	
Exposed	a	—	PY_1
Nonexposed	c	—	PY_0
Total	a + c		T

$$RR = \frac{I_1}{I_0} = \frac{a/PY_1}{c/PY_0}$$

Table 12-7. **Formulas for the calculation of the Mantel-Haenszel** pooled relative risk estimate or its analogues, assuming **uniform stratum-specific estimates**

CASE-CONTROL STUDY:

$$RR_{MH} = \frac{\Sigma ad/T}{\Sigma bc/T}$$

COHORT STUDY WITH COUNT DENOMINATORS:

$$RR_{MH} = \frac{\Sigma a(c+d)/T}{\Sigma c(a+b)/T}$$

COHORT STUDY WITH PERSON-YEARS DENOMINATORS:

$$RR_{MH} = \frac{\Sigma a(PY_0)/T}{\Sigma c(PY_1)/T}$$

tivity and **M** 1, stratified by cigarette smoking (Table 12-5), yields the following:

$$RR_{MH} = \frac{\sum \frac{ad}{T}}{\sum \frac{bc}{T}}$$

$$= \frac{\frac{(41)(52)}{223} + \frac{(41)(39)}{190} + \frac{(22)(26)}{103} + \frac{(86)(40)}{273}}{\frac{(84)(46)}{223} + \frac{(80)(30)}{190} + \frac{(34)(21)}{103} + \frac{(68)(79)}{273}}$$

$$= 0.64$$

This value means that, once differences in cigarette smoking have been taken into account, the relative risk of MI associated with high levels of physical activity is 0.64. This estimate is adjusted for, or unconfounded by, cigarette smoking.

The magnitude of confounding in any study is evaluated by observing the degree of discrepancy between the crude and adjusted estimates. The fact that the crude and adjusted relative risk estimates in the example above are identical indicates that there was no confounding of the association between physical activity and MI by cigarette smoking as categorized in these data.

As discussed earlier in this chapter, the presence or absence of confounding should never be assessed by using a statistical test of significance. A large sample size could easily result in a statistically significant association between the confounder and the exposure or disease, even though the magnitude may be too small to result in any material amount of confounding. On the other hand, even strong associations that could produce confounding of substantial epidemiologic importance may fail to reach statistical significance with a small sample size. Significance testing can be used, however, to evaluate whether the unconfounded **esti-** mate of effect differs from the null value of no association. Hypothesis testing for stratified data is a straightforward extension of the tests applied to crude data. The components of the test statistic are now derived not from a single table, but as a sum of the relevant components in each of the strata. The Mantel-Haenszel test statistic with one degree of freedom is a simple extension of the chi-square formula for a series of **two-** by-two tables of either case-control or cohort data with count denominators. This can be expressed as

$$\chi^2_{MH} = \frac{\left[\sum a - \sum \frac{(a+b)(a+c)}{T}_I\right]^2}{\sum \frac{(a+b)(c+d)(a+c)(b+d)}{T^2(T-1)}}$$

Analogously, the formula for a cohort study with person-year denominators is

$$\chi^2_{MH} = \frac{\left[\sum a - \sum \frac{(a+c)(PY_1)}{T}_I\right]^2}{\sum \frac{(a+c)(PY_1)(PY_0)}{T^2}}$$

Thus, to evaluate the likelihood that the association between physical activity and risk of MI, after adjustment for cigarette smoking, is due to chance, the Mantel-Haenszel chi-square statistic can be calculated as follows:

$$\chi^2_{MH} = \frac{\left[190 - \left(\frac{(125)(87)}{223} + \frac{(121)(71)}{190} + \frac{(56)(43)}{103} + \frac{(154)(165)}{273}\right)\right]^2}{\frac{(125)(98)(87)(134)}{223^2(222)} + \frac{(121)(69)(71)(119)}{190^2(189)} + \frac{(56)(47)(43)(60)}{103^2(102)} + \frac{(154)(119)(165)(108)}{273^2(272)}}$$

$$= 9.16,\ P = 0.0029$$

This means that in these data, there is a statistically significant association between level of physical activity and risk of MI, once the effect of cigarette smoking has been taken into account. It should be noted that in such a circumstance, where the stratified analysis demonstrated no confounding by a given variable, hypothesis testing could have been performed on the crude rather than the stratified data. Again, as with the analysis of crude data, exact confidence intervals can be calculated [8], as can a number of forms of approximate confidence limits [12, 24, 27]. The test-based confidence interval [20] is particularly easy to compute by substituting into the general formula the pooled estimate of the relative risk, as well as the value from the Mantel-Haenszel chi-square statistic. In the example above of physical activity and MI, this calculation would be as follows:

$$95\%\ CI = RR_{,,,}{}^{\pm 1.96/\sqrt{\chi^2_{MH}}}$$

$$= 064^{\pm 1.96/\sqrt{9.16}}$$

$$= (0.48, 0.85)$$

In summary, these data indicate a statistically significant inverse association between level of physical activity and risk of MI once the effect **of cigarette smoking has been taken into account.** The overall estimate of effect is **0.64** and, with 95-percent confidence, the true relative risk **lies between 0.48 and 0.85. Cigarette smoking is not a confounder of the association in these data. Moreover, the magnitude of the observed reduction in risk of MI associated** with increased levels of physical activity **is the same regardless of' whether an individual is a lifelong nonsmoker,** an exsmoker, or a current smoker.

The fact that the stratum-specific estimates were similar or uniform indicates that the magnitude of association between physical activity and **MI did not change according to the level of a third variable, cigarette** smoking. This suggests that there was no modification by cigarette **smoking of the effect of physical activity on risk of MI. In other words, regardless of' a person's smoking history, there was a similar protective effect on risk of MI observed with increased level of physical activity. A different situation exists with respect to the interrelationship between physical activity, MI, and gender. As with cigarette smoking, gender is a potential confounder, since it may be associated with degree of physical** activity and is independently associated with risk of MI. As seen in Table **12-8,** however, when the data are stratified by gender, the association between physical activity and MI appears markedly different for men and women.

Specifically, among men, there is a statistically significant protective effect of physical activity on risk of **MI (RR = 0.53, 95% CI =** 0.38–0.72). In contrast, for women there was no evidence of an inverse association between physical activity and MI, and in **fact,** the data are compatible with the possibility of a small increased risk (RR = 1.19, 95% **CI = 0.65–2.16). There** are a number of possible explanations for this finding, including a true physiologic difference between men and women concerning **the effect of physical activity on risk of MI. The relatively** wide confidence interval, however, suggests that the sample size of women available may simply have been too small to estimate the magnitude of the association in this subgroup with any precision. The important point is that any such possible effect modification of the association between physical activity and MI by gender is best reported and explored. The nature of the relationship between physical activity and MI can be explored in the pattern of the stratum-specific estimates. If only a single summary estimate were reported, the fact that there was effect modification-that is, that the magnitude or, even more extreme, the direction of the association under study differs for varying levels of another factor-would be completely obscured.

When there is effect modification-that is, when the association between the exposure and disease under study varies by levels of a third factor-the emphasis in the analysis of the data and the presentation of

Table 12-8. **Data from a case-control study of physical activity and risk of MI, stratified by gender**

Gender	Physical activity index	Cases	Controls	Total	RR, 95% CI
Men	2500 + kcals	141	208	349	
	< 2500 kcals	144	112	256	0.53, (0.38–0.73)
	Total	285	320	605	
Women	2500 + kcals	49	58	107	
	< 2500 kcals	32	45	77	1.19, (0.65-2.16)
	Total	81	103	184	
Total	2500 + kcals	190	266	456	
	< 2500 kcals	176	157	333	0.64
	Total	366	423	789	

the study results should be on describing how the association of interest is modified by the stratification factor. The first step in any stratified analysis, therefore, is the determination of whether effect modification is present in the data. In most circumstances, this decision should be based on simply "eyeballing" the data to judge the observed patterns of variation. This should be performed in the context of evidence from other investigations to achieve a biologic understanding of the nature of the association under study. If a more formal statistical evaluation of the uniformity of the stratum-specific estimates is desired, a variety of **chi-**square tests for homogeneity are available that test the null hypothesis that the degree of variability in the series of stratum-specific estimates is consistent with random variation **[29].** Again, however, statistical testing to determine the presence or absence of effect modification should only be used as a guide, since statistical significance is so heavily influenced by sample size.

When the stratum-specific estimates vary sufficiently to indicate that there is likely to be variation in the underlying magnitude of the **association** between exposure and disease, the primary and most informative approach to the presentation of the data **is** to **report** separately for each stratum the estimate of effect as well as the confidence interval. The calculation of a summary unconfounded measure is possible but much less important because it **does not represent** the nature of the association observed in all of the groups If, however, it seems desirable to calculate, **in** addition to the stratum-specific estimates, a single unconfounded measure of effect, it would not be appropriate to **use a** pooled estimate since the weights for the pooled estimate have been chosen to provide the most **precise** estimate-of a uniform value of effect. When there is effect modification, specific weights are selected to standardize the **stra-**

Table 12-Y. **Formulas for the calculation of a relative risk estimate standardized to the exposed population, assuming stratum-specific relative risks to be non-uniform**

CASE-CONTROL STUDY:

$$RR_{STD} = \frac{\Sigma a}{\Sigma \frac{bc}{d}}$$

COHORT STUDY WITH COUNT DENOMINATORS:

$$RR_{STD} = \frac{\Sigma a}{\sum \frac{(c)(a + b)}{(c + d)}}$$

COHORT STUDY WITH PERSON-YEARS DENOMINATORS:

$$RR_{STD} = \frac{\Sigma a}{\sum \frac{(c)(PY_1)}{PY_0}}$$

turn-specific values to standard distributions such as those of the exposed or nonexposed populations. As summarized in Table 12-9, formulas for the calculation of the standardized relative risk estimate are available for case-control studies, cohort studies with count denominators, and cohort studies with person-years.

Applying these formulas to the case-control data in Table 12-8 would lead to a standardized relative risk estimate of the association between level of physical activity and MI, adjusted for gender, as follows:

$$RR_{STD} = \frac{\sum a}{\sum \frac{bc}{d}}$$

$$= \frac{190}{\frac{(208)(144)}{112} + \frac{(58)(32)}{45}}$$

$$= 0.62$$

The crude (KK = 0.64) and adjusted (RR = 0.62) relative risk estimates are similar, indicating that, as with cigarette smoking, there was virtually no confounding of the association under study by gender. Unlike cigarette smoking, however, gender is a possible effect modifier of the association between physical activity and risk of MI.

For any association under study, a given factor can be both a confounder and an effect modifier, a confounder but not an effect modifier, an effect modifier but not a confounder, or neither. Confounding and effect modification are very different in both the information each provides and what is done with that information. Whether a factor is a confounder or not depends solely on whether it is distributed unevenly between the study groups and may thus in itself have accounted, totally or in part, for the observed association between exposure and disease. Confounding is a nuisance effect, resulting in a distortion of the true relationship between the exposure and risk of disease due solely to the particular mix of subjects included in the study. In another investigation, with a different group of subjects or a different design, the same variable may not be a confounder of the same association [16]. Thus, the aim is to control confounding and eliminate its effects. Effect modification is assessed by determining whether the magnitude or even direction of the association under study varies according to the presence or level of a third factor. This reflects a characteristic of nature that exists independently of any particular study design or subjects. Whether the association between physical activity and MI is different for men and women (effect modification) is in no way influenced by whether men and women in the study sample are likely to have different physical activity levels (confounding). Effect modification answers the question of whether the relationship between the exposure and disease appears to be the same or different for varying levels of a factor after baseline differences in that factor are controlled. Effect modification is to be described and reported, not controlled. Exploring the nature of the interaction, whether qualitative or quantitative, multiplicative or additive, may provide tremendous insight into the interrelationships between, and mechanisms of action of, the exposures and disease [13, 23, 26].

The process of stratification is used to evaluate both confounding and effect modification, to control the former, and to describe the latter. As summarized in Table 12-10, the approach to a stratified analysis involves a number of steps. First, the data representing the overall crude association between exposure and disease is stratified by levels of the potential confounding factor. Stratum-specific relative risk estimates are then calculated, each of which is unconfounded. To decide how best to present the data, the stratum-specific relative risk estimates are then evaluated for similarity, either by "eyeballing" or by performing a test of statistical significance. If the effects are thought to be uniform, a single summary unconfounded relative risk estimate can be calculated by pooling the stratum-specific values, using a Mantel-Haenszel relative risk estimator. Hypothesis testing is then performed on the unconfounded estimate using the appropriate Mantel-Haenszel chi-square, and the confidence interval is computed. If the stratum-specific estimates are not uniform, effect modification may be present, and the stratum-specific values should be reported and described. In addition, a single summary esti-

Table 12-10. **Steps for the control of confounding and the evaluation of effect modification through stratified analysis**

1. Stratify by levels of the potential confounding factor.
2. Compute stratum-specific unconfounded relative risk estimates.
3. Evaluate similarity of the stratum-specific estimates by either eyeballing or performing test of statistical significance.
4. If effect is thought to be uniform, calculate a pooled unconfounded summary estimate using RR_{MH}.
5. Perform hypothesis testing on theunconfounded estimate, using Mantel-Haenszel chi-square and compute confidence interval.
6. If effect is not thought to be uniform (i.e., if effect modification is present):
   a. Report stratum-specific estimates, results of hypothesis testing, and confidence intervals for each estimate
   b. If desired, calculate a summary unconfounded estimate using a standardized formula.

mate that is unconfounded can be calculated using a standardized measure.

Mu&variate Analysis

As illustrated in the previous section, stratification in the analysis as a technique to control confounding has a number of advantages. It is easy to carry out; it permits the evaluation of effect modification; and, perhaps more importantly, it allows both the investigators and the readers to achieve a clear understanding of the interrelationships among the exposure, disease, and additional confounding and effect modifying variables. A fundamental problem with stratified analysis, however, is its inability to control simultaneously for even a moderate number of potential confounders. For example, in the study of physical activity and MI, suppose that the investigators chose to consider only four potential confounding factors: sex (male, female), age (less than 50, 50-59, 60-69, 70 +), smoking status (never smoker, past smoker, current smoker), and Quetelet's index, a measure of obesity defined as weight in pounds divided by height in inches squared (lowest, second, third, highest quartiles). These variables would require a total of $2 \times 4 \times 3 \times 4 = 96$ strata to represent the possible combinations of sex, age, smoking, and obesity. Even with a relatively large study, it is very likely that many of these **strata** would contain few, if any, individuals, making analysis unreliable or even impossible. The control of an additional variable would necessitate multiplying the 96 strata by the number of values of still another factor, and the total number of individuals in the sample would very quickly become inadequate.

Multivariate analysis allows for the efficient estimation of measures of association while controlling for a number of confounding factors simultaneously, even in situations where stratification would fail because of insufficient numbers. In general, a multivariate technique refers to any analysis of data that takes into account a number of variables simultaneously. All types of multivariate analysis involve the construction of a mathematical model to describe most efficiently the association between exposure and disease as well as other variables that may confound or modify the effect of exposure. A large number of multivariate models have been developed for specialized purposes, each with a particular set of assumptions underlying its applicability. The choice of the appropriate model is complex and is based on the underlying design of the study, the nature of the variables, as well as assumptions regarding the interrelationship between the exposures and outcomes under investigation. Multivariate analysis is the subject of many advanced textbooks [1, 2, 5, 11, 12]. The application of the techniques, especially the selection and form of the categorization of the variables for inclusion into the model, requires expert guidance. It is beyond the scope of this textbook to provide a full discussion of each multivariate technique and its underlying theory and applications. What we wish to do here, however, is to describe intuitively the rationale for the use of multivariate analysis in epidemiologic research, a few of the strategies that are most often encountered, and a brief discussion of their strengths and limitations.

The most common way that many factors are controlled for simultaneously is through the use of a multiple regression model. Multiple regression is an extension of the most fundamental model describing the relationship between two variables, namely a straight line. In simple linear regression, the relationship between the mean of the dependent, or outcome, variable (Y) and an independent, or predictor, variable (X) can be simply expressed as $Y = a + bX$. In this equation, a is the linear regression intercept or constant, that is, the average value of the dependent variable when X = 0, and the coefficient b is the slope of the line representing the association between X and Y, that is, the estimated mean change in the expected value of the dependent variable for each unit change of the independent variable. For example, in evaluating the relationship between the independent variable age and systolic blood pressure, the coefficient b would indicate the mean change in systolic blood pressure for every unit change (e.g., year) of age. Multiple linear regression involves expanding this equation to include a number of independent variables:

$$Y = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$$

where:

n = the numher of independent predictor variables

$X_1 \ldots X_n$ = an individual's particular set of values for the independent variables

$b_1 \ldots b_n$ = the respective coefficients for each of the independent variables

As with every type of mathematical model, linear regression involves assumptions about the characteristics of and relationship between the dependent and independent variables. Specifically, both simple and multiple regression assume a linear function of the variables included in the model. While few relationships between variables are strictly linear, often the true association is close enough to allow reasonable inferences to be made from such a model. The adequacy of a linear fit can be tested by a number of statistical techniques [4, 14]. If the relationship between variables is inherently nonlinear, some type of transformation can often be performed to accommodate the situation. For example, the incidence of many diseases, such as cancer, increases with age, but as a curved or exponential rather than linear relationship. A linear regression that utilized the square of age rather than age itself would result in a better representation of the true association between age and incidence of disease and enable a linear model to accommodate a basically nonlinear relationship.

The coefficients in the model result from a method that calculates the "best fitting" line, that is, the unique line that best describes the population that gave rise to the observed data. The common method of calculating these coefficients is that of least squares, which minimizes the sum of the squared deviations of each observation from the fitted line. The coefficient of each independent variable can be interpreted as the magtude of the increase (or, if negative, the decrease) in the value of the mean of the dependent variable for every unit increase in that predictor variable, caking into account the effect of all other variables in the model. For example, with respect to evaluating the relationship of age and systolic blood pressure, a number of potential confounders such as sex, obesity, smoking, exercise, and dietary factors could also be entered into the model. The coefficient for age in such a model would indicate the mean change in systolic blood pressure for every unit change in age, controlling for the effects of all these other variables. It is just this conditional interpretation of the coefficients of multivariate analysis that makes this analytic technique so useful for epidemiologic studies, because the effect of any variable unconfounded by all other factors in the model can be assessed.

Multiple linear regression is most appropriately used when the dependent outcome variable is continuously distributed, as with levels of blood pressure or serum cholesterol. In many epidemiologic studies, the outcome of interest is a binary variable such as diseased versus nondiseased or dead versus alive. In such circumstances, it is possible to use a specialized type of multiple regression called logistic regression analysis, which is a powerful statistical tool for estimating the magnitude of the association between an exposure and a binary outcome after adjusting simultaneously for a number of potential confounding factors. This model is a simple variant of the multiple regression equation, in which the risk of developing an outcome is expressed as a function of independent predictor variables. Specifically, the dependent variable is defined as the natural logarithm (ln) of the odds of disease, or the logit. If Y is the probability of disease, then $Y/(1 - Y)$ represents the "odds" of developing the outcome, and the log odds of disease, or the logit, can be written as $\ln[Y/(1 - Y)]$. The log odds of disease as the dependent variable can then be expressed as a simple linear function of the independent predictor variables using the following formula:

$$\ln\left[\frac{Y}{1-Y}\right] = a + b_1X_1 + \ldots + b_nX_n$$

This equation can be rewritten to represent the probability of disease as:

$$Y = \frac{1}{1 + e^{-(a + b_1X_1 + \ldots + b_nX_n)}}$$

As with multiple linear regression, the specific set of values for the intercept a and for $b_1, \ldots, b_n$ in logistic regression are calculated to represent those that provide the most likely estimate of the population from which the observed data arose. Since the logistic model for the probability of disease yields values that are always between zero and one, its use has important implications for the interpretation of the coefficients. The coefficients obtained through logistic regression by definition describe increases or decreases in the log odds produced by one unit of change in the value of the independent variable and thus indicate the effect of an individual factor on the log odds of the outcome controlling for the effects of all remaining variables held constant. A practical advantage of logistic regression over multiple linear regression in epidemiologic research is that these coefficients can be directly converted to an odds ratio that provides an estimate of the relative risk that is adjusted for confounding. If the independent variable is binary, the antilogarithm of the coefficient for that variable in a logistic regression is the odds ratio representing the magnitude of the association between the factor and

the **outcome,** controlling for the effects of all other variables in the model [30]:

$$RR\ (x_i) = e^{b_i}$$

Confidence limits around this estimate of relative risk can then be obtained using the coefficient and its related **standard** error:

$$95\%\ CI = e^{(b_i \pm 1.96SE_{b_i})}$$

Relative risk estimates for increments of independent variables that are ordinal or continuous can also be obtained [12, 25]. Alternatively, such variables can be redefined and entered into the model as a series of binary variables, and the above formula can be used to calculate the odds ratio of developing the disease at a given level of exposure relative to the specified referent group.

The use and interpretation of logistic regression can be illustrated with the previous example of level of physical activity and risk of MI. Because the univariate stratified analyses suggested possible effect modification of the association by gender, separate multivariate analyses were performed for men and women. Variables entered into the model included demographic factors (age, educational level), medical history (diastolic blood pressure, treatment of hypertension, family history of MI, obesity, diabetes, cholesterol level), and life-style factors (cigarette smoking, personality type, alcohol consumption, total daily calories, daily saturated fat intake). The coefficient associated with a Physical Activity Index of 2500 **kcal** per day, controlling for the simultaneous effects of the above variables, was -0.766, with a standard error of 0.266. Thus, the relative risk estimate for physical activity and MI among men, controlling for confounding by these factors, is as follows:

$$RR = e^{(-0.766)}$$
$$= 0.46$$

This compares with the calculated crude value of 0.53, indicating that there was little confounding of the association between physical activity and MI by the combined effect of these factors. The **95-percent** confidence interval around the adjusted estimate can be calculated as

$$95\%\ CI = e^{[-0.766 \pm 1.96(0.266)]}$$
$$= (0.28,\ 0.78)$$

This model for logistic regression assumes that the estimate of effect is uniform across the different levels of the confounding variables, that

is, there is no effect modification by the variables included in the model. If that assumption cannot be made, a product term representing a combination of the exposure and potential effect modifier can also be added to the model to allow for the assessment of interactions. Thus, to evaluate, for example, possible effect modification of the association among men between physical activity and MI by family history of the disease, the logistic regression model could be set up as follows:

$$\ln\left[\frac{Y}{1-Y}\right] = a + b_1X_1 + b_2X_2 + b_3(X_1X_2) + \ldots + b_nX_n$$

where:

Y = probability **of MI**

X_1 = physical activity level (I = high, 0 = low)

X_2 = family **history of MI (1 = present, 0 = absent)**

X_1X_2 = **status on both physical activity and family history of MI**

$x_3 \ldots X_n$ = **other independent variables**

$b_1 \ldots b_n$ = **the respective coefficients of each of the independent variables or combination** of **variables**

The coefficient of the product term (b_3) permits assessment of whether the magnitude of the contribution of physical activity to the log odds of MI varies according to family history of the disease, that is, whether there is evidence of effect modification. The calculation of the effect of physical activity on risk of MI will then involve information from b_1, the coefficient for physical activity, as well as b_3 [12].

In studies where participants have not all been followed for an equal period of time to determine the development of the outcomes of interest, the multivariate techniques that have been described are not applicable. A specialized type of multivariate analysis, called the proportional hazards model, was developed by Cox [5] to take into account the unequal lengths of time that each cohort member is observed for the occurrence of the outcome of interest. The proportional hazards model describes the relation of the independent variable to the natural logarithm of the incidence rate of disease rather than to the odds of disease. The form of this model is similar to that of multiple logistic regression but slightly more complicated since the components must be dependent on time. Specifically:

$$\text{In [incidence rate (t)]} = a(t) + b_1X_1 + \ldots + b_nX_n$$

where:

$a(t)$ = the **baseline incidence rate expressed as a function of time**

$X_1, \ldots X_n$ = **a person's particular set of values for the independent variables**

$b_1, \ldots b_n$ = the respective coefficients for each independent variable

The construction and interpretation of coefficients from the proportional hazards model as estimates of the incidence density ratio is analogous to that of the coefficients from the logistic regression model.

Multivariate modeling can be used in epidemiologic research for both explanatory and predictive purposes. With respect to explanatory uses, as in the examples above, the interrelationships between the exposure, disease, and other factors can be described. With respect to predictive purposes, the estimated coefficients can be used to calculate the probability that an individual having a specific set of values for the predictor variables will experience the outcome of interest. This can be accomplished in a straightforward manner in a cohort study. In a case-control study, on the other hand, because the relative numbers of cases and controls in the sample are determined by the investigator and not by the incidence of disease in the population, the intercept term for the prediction equation is simply an artifact of the study design. Thus, as with case-control studies in general, it is not possible using multivariate modeling to calculate absolute rates or risks for an individual with a particular set of values for the predictor variables unless information on the absolute rate of disease is available from sources outside the study. The study results can, however, provide the relative magnitudes of the rates of disease for subjects having various sets of values for the predictor variables.

Perhaps the main disadvantage of all multivariate techniques is that the process of efficient mathematical modeling can often occur at the expense of a clear understanding of the data by either the investigator or the reader. In many ways, the use of multivariate analysis can appear like a "black box" strategy, in which all of the variables are entered into a specialized computer program, and the net result is a single value representing the magnitude of the association between the exposure and disease after the effects of all confounders have been taken into account. As discussed in the beginning of this section, the selection of the appropriate model and variables, as well as the way in which these variables are entered into the model, is a complex issue that should be accomplished in consultation with someone experienced in the use of these techniques. It can be difficult for the investigator to both understand and communicate the results of this process to the reader, as well as to keep clearly in mind the limitations of the data from which the results arose. Thus, while multivariate analysis is undeniably useful for the control of confounding when stratified analysis is impractical, it is important that it be performed and interpreted carefully and that the control of confounding by this method not be achieved at the expense of lack of familiarity with the basic data. A stratified analysis should always be examined prior to performing multivariate procedures and presented in conjunction with the multivariate results. This will help ensure an understanding of the nature of the individual confounding variables as well as the interactions between the factors under study, assist in choosing the particular variables to be included in the multivariate analysis, and provide information concerning the degree to which the data conform to the assumptions of the particular multivariate model chosen. This thorough exploration of the basic data prior to the more complex analyses can be easily accomplished with available programs suitable for hand-held programmable calculators [25] and personal computers.

CONCLUSION

In all analytic studies, particularly observational case-control and cohort designs, confounding must always be considered as an alternative explanation for study findings. There are a number of methods available for the control of confounding in the design or analysis of any study. These include restriction, matching, or randomization (in clinical trials) in the design as well as stratification and multivariate techniques in the analysis. No single method can be considered optimal in every situation. Each has its strengths and limitations, which must be carefully considered at the beginning of the study. In most situations, a combination of strategies will provide better insights into the nature of the data and more efficient control of confounding than any single approach.

STUDY QUESTIONS

1. In matched-pair designs, why is it always necessary first to conduct a matched analysis? In what circumstances can the matching safely be disregarded?
2. In stratified analyses, compare and contrast the evaluation of confounding and effect modification.
3. What are the chief strengths and limitations of multivariate analysis?

REFERENCES

1. **Bishop, Y. M. M.,** Fienberg, S. E., **and Holland, P. W.** *Discrete Multivariate Analysis: Theory and Practice.* **Cambridge, MA: MIT Press, 1975.**
2. **Breslow, N. E., and Day,** *N. E. Statistical Methods in Cancer Research: Vol. I. The Analysis of Case-Control Studies.* **Lyon, France: IARC Scientific Publications, 1981.**

3. Buring, J. E., Willett, W., Goldhaber, S. Z., et al. Alcohol and HDL in non-fatal myocardial infarction: Preliminary results from a case-control study. *Circulation* 68:227, 1983.
4. Colton, T. *Statistics in Medicine.* Boston: Little, Brown, 1974.
5. Cox, D. R. *Analysis of Binary Data.* London: Methuen, 1970.
6. Doll, R., and Hill, A. B. Smoking and carcinoma of the lung: Preliminary report. Hr. *Med. J.* 2:739–748, 1950.
7. Fleiss, J. L. *Statistical Methods* for *Rates and Proportions. New* York: Wiley-Inter-science, 1973.
8. Gart, J. J. Point and interval estimation of the common odds ratio in the combination of 2 x 2 tables with fixed margins. *Biometrika* 57:471-475, 1970.
9. Greenland, S., and Robins, J. M. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 41:55, 1985.
10. Hennekens, C. H. Alcohol. In N. Kaplan and J. Stamler (eds.), *Prevention of Coronary Heart Disease.* Philadelphia: Saunders, 1983. Pp. 130-138.
11. Kleinbaum. D. G., and Kupper, L.. L,. *Applied Regression Analysis and Other Multivariable Methods.* Boston: Duxbury Press, 1978.
12. Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, CA: Lifetime Learning Publications, 1982.
13. Kupper, L., and Hogan, M. D. Interaction in epidemiologic studies. *Am. J. Epidemiol.* 108:447–453, 1978.
14. Lemenstrow, S., and Hosmer, D. W. A review of goodness of fit statistics for use in development of logistic regression models. *Am. J. Epidemiol.* 115:92–106, 1982.
15. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J.N.C.I.* 22:719–748, 1959.
16. Miettinen, 0. S. Confounding and effect modification. *Am. J. Epidemiol.* 100:350, 1974.
17. Miettinen, O. S. Matching and design efficiency in retrospective studies. *Am. J. Epidemiol.* 91:111, 1970.
18. Miettinen, 0. S. Individual matching with multiple controls in the case of all or none responses. *Biometrics* 22:339–355, 1969.
19. Miettinen, O. S. Estimation of relative risk from individually matched series. *Biometrics* 26:75–86, 1970.
20. Miettinen, 0. S. Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* 103:226–235, 1976.
21. Prentice, R. Use of the logistic model in retrospective studies. *Biometrics* 32:599–606, 1976.
22. Rosner, B., and Hennekens, C. H. Analytic methods in matched pair epidemiologic studies. *Int. J. Epidemiol.* 7:367–372, 1978.
23. Rothman, K. J. Synergy and antagonism in cause-effect relationships. *Am. J. Epidemiol.* 103:506–511, 1976.
24. Rothman, K. J. *Modern Epidemiology.* Boston: Little, Brown, 1986.
25. Rothman, K. J.. and Boice, J. D., Jr. *Epidemiologic* Analysis *with* a *Programmable Calculator.* U.S. D.H.E.W. N.I.H. Publication No. 79-1649, 1979.
26. Rothman, K. J., Greenland, S., and Walker, A. M. K. Concepts of interaction. *Am. J. Epidemiol.* 112:467–470, 1980.

27. Schlesselman, J. J. *Case-Control Studies: Design, Conduct, Analysis. New* York: Oxford University Press, 1982.
28. Smith, D. C., Prentice, R., Thompson, D. J., et al. Association of exogenous estrogen and endometrial carcinoma. N. *Engl. J. Med. 293:* 1164-1167, 1975.
29. Snedecor, G. W., and Cochrane, W. C. *Statistical Methods* (6th ed.). Ames, IA: Iowa State University Press, 1967.
30. Truett, J., Cornfield, J., and Kannel. W. A multivariate analysis of the rick of coronary heart disease in Framingham. *J. Chronic Dis.* 20:511-524, 1967.