

3. GENERAL CONSIDERATIONS FOR THE ANALYSIS OF CASE-CONTROL STUDIES

- 3.1 Bias, confounding and causality
- 3.2 Criteria for assessing causality
- 3.3 Initial treatment of the data
- 3.4 Confounding
- 3.5 Interaction and effect modification
- 3.6 Modelling risk
- 3.7 Comparisons between more than two groups
- 3.8 Considerations affecting interpretation of the analysis

CHAPTER III

GENERAL CONSIDERATIONS FOR THE ANALYSIS OF CASE-CONTROL STUDIES

In previous chapters we have introduced disease incidence as the basic measure of disease risk in a population. As a measure of the increased risk for a population exposed to some factor when compared with an otherwise similar population not so exposed, we have proposed the use of the proportionate increase in incidence which corresponds to the relative risk. We have described the properties of this measure, and its behaviour in cancer epidemiology, in order to demonstrate its advantages over an alternative measure of disease association, the excess risk. We have explored the logical basis for estimation of the relative risk from the results of a case-control study, from which the actual incidence rates cannot be estimated. Estimation of relative risks follows from interpreting the case-control study as the result of sampling from a large, probably fictive, cohort study from which incidence rates can hypothetically be estimated. In succeeding chapters we shall develop the statistical theory and methodology required for the analysis of case-control data. In this chapter, we shall concern ourselves with the types of conclusion that we want to draw from the data, and the steps which must be taken to ensure that these conclusions are valid. Strategies for approaching the data, the handling of different types of variables, the examination of joint association of several variables, and how the design of a study is reflected in the analysis will all be discussed.

3.1 Bias, confounding and causality

The purpose of an analysis of a case-control study is to identify those factors under study which are associated with risk for the disease. In an analysis, the basic questions to consider are the degree of association between risk for disease and the factors under study, the extent to which the observed associations may result from bias, confounding and/or chance, and the extent to which they may be described as causal. The concepts of bias and confounding are most easily understood in the context of cohort studies, and how case-control studies relate to them. Confounding is intimately connected to the concept of causality. In a cohort study, if some exposure E is associated with disease status, then the incidence of the disease varies among the strata defined by different levels of E . If these differences in incidence are caused (partially) by some other factor C , then we say that C has (partially) confounded the association between E and disease. If C is not causally related to disease, then the differences in incidence cannot be caused by C , thus C does not confound the disease/exposure association. Often the observed extraneous variables will only be surrogates for the factor causally related to disease,

age and socioeconomic status being obvious examples, but we should normally consider these surrogates also as confounding variables. Confounding in a case-control study has the same basis as in a cohort study. It arises from the association in the causal network in the underlying study population and cannot normally be removed by appropriate study design alone. An essential part of the analysis is an examination of possible confounding effects and how they may be controlled. Succeeding chapters consider this problem in detail.

Bias in a case-control study, by contrast, arises from the differences in design between case-control and cohort studies. In a cohort study, information is obtained on exposures before disease status is determined, and all cases of disease arising in a given time period should be ascertained. Information on exposure from cases and controls is therefore comparable, and unbiased estimates of the incidence rates in the different subpopulations can be constructed. In case-control studies, however, information on exposure is normally obtained after disease status is established, and the cases and controls represent samples from the total. Biased estimates of incidence ratios will result if the selection processes leading to inclusion of cases and controls in the study are different (selection bias) or if exposure information is not obtained in a comparable manner from the two groups, for example because of differences in response to a questionnaire (recall bias). Bias is thus a consequence of the study design, and the design should be directed towards eliminating it. The effects of bias are often difficult to control in the analysis, although they will sometimes resemble confounding effects and can be treated accordingly (see § 3.8).

To summarize, confounding reflects the causal association between variables in the population under study, and will manifest itself similarly in both cohort and case-control studies. Bias, by contrast, is not a property of the underlying population and should not arise in cohort studies. It results from inadequacies in the design of case-control studies, either in the selection of cases or controls or from the manner in which the data are acquired.

It is not helpful to introduce the concepts of necessity or sufficiency into the discussion of causality in cancer epidemiology. Apart from occasional extremes of occupational exposure, constellations of factors have not been identified whose presence inevitably produces a cancer, or, conversely, in whose absence a tumour will inevitably not appear. Thus, we shall use the word "cause" in a probabilistic sense. By saying that a factor is a cause of a disease, we mean simply that an increase in risk results from the presence of that factor. From this viewpoint a disease can have many causes, some of which may operate synergistically. It is sometimes helpful to think in terms of a multistage model, and to consider a cause as a factor which directly increases one or more of the rates of transition from one stage to the next (Peto, 1977; Whittemore, 1977a). One factor may need the presence of another to be effective, in which case one should strictly speak of the joint occurrence as being a cause.

The most one can hope to show, even with several studies, is that an apparent association cannot be explained either by design bias or by confounding effects of other known risk factors. There are, nevertheless, several aspects of the data, even from a single study, which would make one suspect that an association is causal, which we shall now discuss (Cornfield et al., 1959; Report of the Surgeon General, 1964; Hill, 1965).

3.2 Criteria for assessing causality

Dose response

One would expect the strength of a genuine association to increase both with increasing level of exposure and with increasing duration of exposure. Demonstration of a dose response is an important indication of causality, while the lack of a dose response argues against causality.

In Chapter 2, we saw several examples of a dose response. Table 2.8 shows a smooth increase in risk for oral cancer with increasing consumption of both alcohol and tobacco, and Table 2.7 displays the increasing risk for breast cancer with increasing age at birth of first child. The latter example is not exactly one of a dose response since the dose is not defined, but the hypothesis that later age at first birth increases the risk of developing breast cancer is given strong support by the smooth trend.

The opposite situation is illustrated by the association between coffee drinking and cancer of the lower urinary tract. Table 3.1 is taken from a study by Simon, Yen and Cole (1975). Three previous studies (Cole, 1971; Fraumeni, Scotto & Dunham, 1971; Bross & Tidings, 1973) had also shown a weak association between lower urinary tract cancer and coffee drinking, but with no dose response. The authors of the 1975 paper concluded that, taking the four studies together, the association was probably not causal. The three arguments they advanced were: (1) the absence of association in some groups, (2) the general weakness of the association, and (3) the consistent absence of a dose response; the last point was considered the most telling.

A clear example of risk increasing with duration of exposure is given by studies relating use of oestrogens to palliate menopausal symptoms with an increased risk for endometrial cancer (see Table 5.1).

Table 3.1 Association between coffee drinking and tumours of the lower urinary tract^a

Cups of coffee/day	Cases	Controls	Relative risk
<1	10	56	1.0
1-2	74	187	2.2
3-4	30	91	1.9
5+	20	48	2.3

^a Data taken from Simon, Yen and Cole (1975)

Specificity of risk to disease subgroups

Demonstration that an association is confined to specific subcategories of disease can be persuasive evidence of causality, as indicated by the following examples.

In earlier days, when the role of cigarette smoking in the induction of lung cancer was still being established, a persuasive aspect of the data was the finding that when a non-smoker developed lung cancer, it was often the relatively rare adenocarcinoma

Table 3.2 Histological types of lung cancer found in Singapore Chinese females, 1968–73, as related to smoking history^a

Histological type	Number	% smokers
Epidermoid carcinoma	12	83.3
Small-cell carcinoma	13	84.6
Adenocarcinoma	25	28.0
Large-cell carcinoma	8	50.0
Other types	6	50.0
Controls	156	31.4

^a Adapted from MacLennan et al. (1977)

(Doll, 1969). This feature of the disease is shown in data from Singapore in Table 3.2 (MacLennan et al., 1977).

As a second example, the association between benzene exposure and leukaemia is restricted to particular cell types, i.e., acute non-lymphocytic (Infante, Rivisky & Wagoner, 1977; Aksoy, Erdem & Dinçol, 1974). The specificity of the association is perhaps the major reason for regarding it as causal.

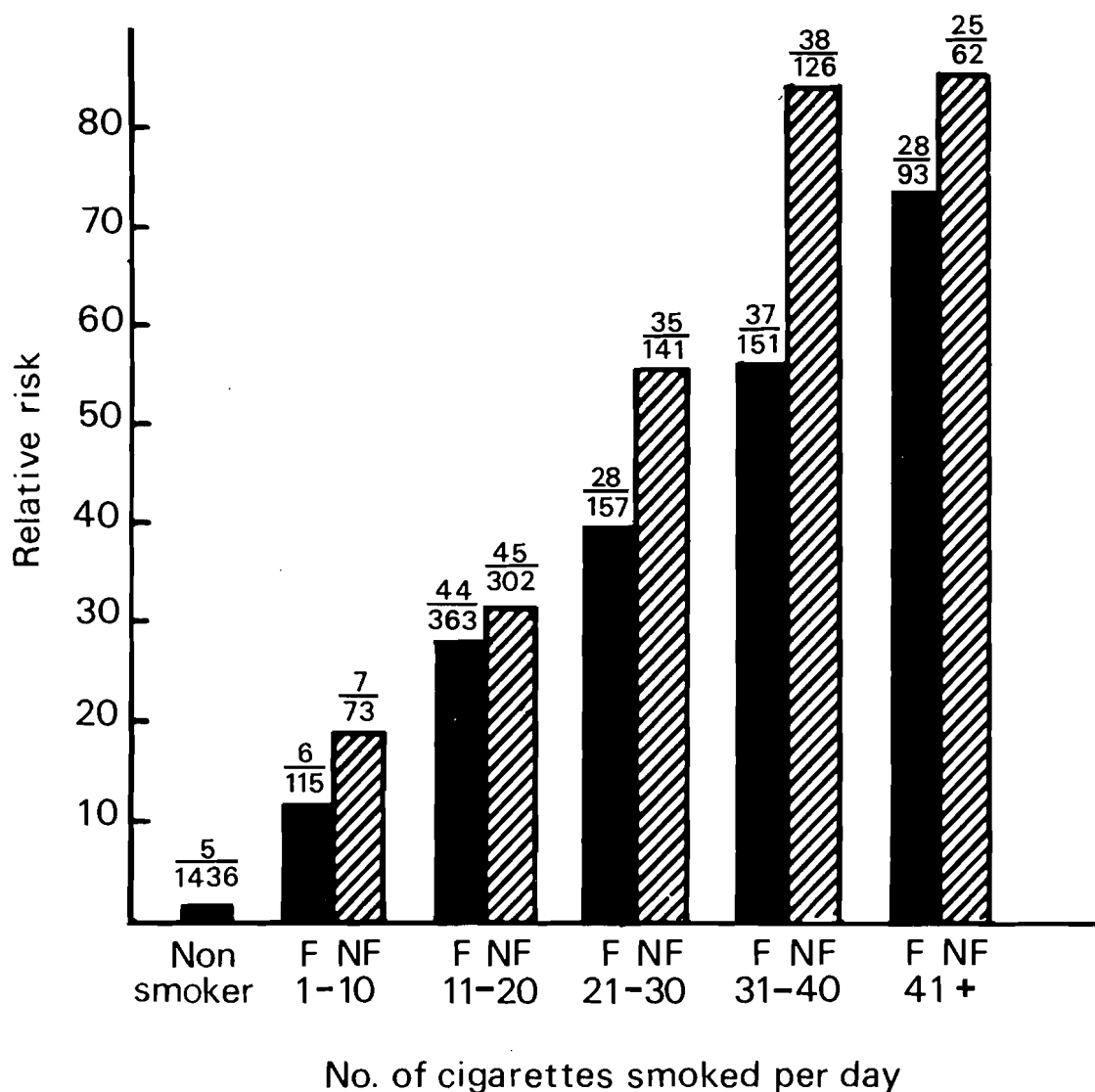
The tendency for several types of cancer to aggregate in families is often difficult to interpret since family members share in part both their environment and their genes. Relative risks for first degree relatives are typically of the order of two- to threefold. The greatly increased familial risk for bilateral breast cancer especially among premenopausal women (Anderson, 1974) reduces the chance that the association is a reflection of either environmental confounding factors or bias in case ascertainment, and enhances one's belief in a genetic interpretation.

Specificity of risk to exposure subcategories

Belief in the causality of an association is also enhanced if one can demonstrate that the disease/exposure association is stronger either for different types of exposures, or for different categories of individuals. A dose response, with higher risk among the more heavily exposed, is an obvious example. Interactions can also provide insight into disease mechanisms. As an example, one can cite the risk for breast cancer following exposure to ionizing radiation: a greater risk was observed for women under age 20 at irradiation than for women irradiated at over 30 years of age (Boice & Monson, 1977; McGregor et al., 1977). Subsequent studies showed that risk was in fact greatly elevated among girls irradiated either in the two years preceding menarche or during their first pregnancy (Boice & Stone, 1978); breast tissue is proliferating rapidly at both these periods of a woman's life.

In Figure 3.1, we show the risk for lung cancer among males associated with smoking varying numbers of filter and non-filter cigarettes. There is a considerably lower risk associated with the use of filter cigarettes, indicating the importance of tars as the carcinogenic constituent of the smoke, since volatile components were not significantly reduced by filters in use at that time (Wynder & Stellman, 1979).

Fig. 3.1 Relative risk for cancer of the lung according to the number of nonfilter (NF) or filter (F) cigarettes smoked per day. Number of cases and controls shown above each bar. From Wynder and Stellman (1979).



Strength of association

Demonstration of a dose response and of variation in risk according to particular exposure or disease subcategories have in common the identification of subgroups at higher risk. In general terms, the closer the association, the more likely one is to consider the association causal. One reason follows directly from a property of the relative risk described in § 2.7. If an observed association is not causal, but simply the reflection of a causal association between some other factor and disease, then this latter factor must be more strongly related to disease (in terms of relative risk) than is the former factor. The higher the risk, the less one would consider that other factors were likely to be responsible. One also has the possibility in all case-control studies that patient

selection or choice of the control group may introduce bias. Bias becomes less tenable as an explanation of an observed association the stronger the association becomes. An example of this is found in the original report on the role of diethylstilboestrol administered to mothers during pregnancy in the development of vaginal adenocarcinomas in the daughters (Herbst, Ulfelder & Poskanzer, 1971). The study was based on 8 cases each with 4 matched controls; 7 out of the 8 cases had been exposed *in utero* to diethylstilboestrol, in contrast to none of the 32 controls. The magnitude of this association persuades one of its causal nature, even though recall of drug treatment some 20 years previously is a potential source of serious bias.

Temporal relation of risk to exposure

For most epithelial tumours, one expects a latent period of at least 15 years. Typically, when exposure is continuous, there is little risk until some 10–15 years after exposure starts, the relative risk then increasing to reach a plateau after 30 years or more (Whittemore, 1977b). For radiation-induced leukaemia this risk increases more quickly (Smith & Doll, 1978), and among recipients of organ transplants the risk for some lymphomas can increase strongly within a year (Hoover & Fraumeni, 1973). Although in principle both cohort and case-control studies should demonstrate the same evolution of relative risk, in practice the temporal evolution of risk following exposure has played a greater role in assessing causality in cohort studies. The reason lies in the nature of the observations. In cohort studies, it is precisely the increase in risk in the years after exposure starts that one observes. Referring back to the discussion of lung cancer risk among smokers in Chapter 2, a prospective study leads to a description of evolution of risk as shown in Figure 2.9 and Table 2.6, whereas a case-control study gives only the relative risk shown in Table 2.6, with most cases probably over age 50. The evolution of risk over time, clear from the changes in the absolute risk in Figure 2.9, is less distinct when considering only the relative risk.

More attention to this aspect of case-control study data may well prove beneficial.

Lack of alternative explanations

In the data being analysed, association between exposures of interest and disease must be shown not to be the effect of some further factor which is itself causally associated with both disease and the exposure. Treatment of potential confounding variables is discussed at length in § 3.4.

Spurious associations can also arise from biased selection of cases or controls, or from biased acquisition of information from either group. Questions of bias are usually more difficult to resolve by considerations internal to the actual data than are problems of confounding. However, if several control groups have been chosen (see § 3.7) or if the data were acquired in a manner in which disease status could not have intervened, the extent to which bias might provide an explanation of the observations is usually reduced.

Considerations external to the study

Magnitude and specificity of risk, dose response and the inability to find alternative explanations are criteria which can be satisfied at least partially by adequate treatment of the data from a single study, and analyses should be aimed in this direction. Comparison can also be made with the trends in the general population, both in terms of the exposure under study and the tumour experience. The early case-control studies of lung cancer were instigated by the parallel increase in cigarette smoking and incidence of the disease. In the paper by Jick et al. (1979) on endometrial cancer, figures are given showing the rise and fall of oestrogen use in the general population and the corresponding rise and fall in the incidence of the disease, together with data showing the high risk among long-term users and the great reduction in risk for individuals who stop taking oestrogens. Arguments offering explanations other than causality for these results would have to be unusually tortuous.

It is rare, however, for a single study to provide convincing evidence of causality. Other studies performed in different populations and using different methodologies are normally required. Demonstration of a reduction in risk after exposure has terminated is further persuasive evidence, although the absence of a reduction is no indication of lack of causality, as asbestos exposure exemplifies (Seidman, Lilis & Selikoff, 1977). Biological plausibility or the demonstration of carcinogenicity in the laboratory provide additional evidence.

General acceptance of the causal nature of an association normally would result only if these more general criteria were satisfied, with several corroborating studies and demonstrations of plausible biological pathways. Nevertheless, even if the results of a single study seldom furnish conclusive evidence of causality, the aim of the analysis should be to extract the fullest evidence for or against causality that the study can provide.

3.3 Initial treatment of the data

The first step in any analysis will be a description of the distribution among cases and among controls of the different variables included in the study. This description should include the correlations, or some other measure of association, between the exposure variables of interest. Such correlations are best computed separately for cases and controls. One would also expect to see a description of the cases and controls in terms of age, sex, and such factors as race, country of birth, hospital attended and method of diagnosis, which although not the object of the study, provide the setting for the interpretation of the later results. It must not be overlooked that the results refer to the sample studied, and generalization from these results usually depends on non-statistical arguments.

Information on exposures which are considered of importance for the cancer site under investigation will usually consist of more than a single measure. For cigarette smoking, for example, one would normally obtain information not only on the daily consumption of cigarettes, but also the age at which smoking started, and stopped if the individual no longer smokes. One may be tempted to proceed directly to a compos-

ite measure, such as cumulative exposure; such a procedure, however, may obscure important features of the disease/exposure association.

For continuing smokers, with data on cigarette consumption obtained retrospectively, one might expect lung cancer incidence to be proportional to the fourth power of duration of smoking, but related linearly to the average daily consumption of cigarettes (Doll, 1971). A man aged 60 who has smoked 20 cigarettes a day since age 40 will have one eighth the risk of a man of the same age who has smoked 10 cigarettes a day since age 20. The total cigarette consumption is the same in the two cases, but the difference in risk is eightfold. Similar differences will be seen if one considers ex-smokers. Twenty years after stopping smoking, the lung cancer risk is approximately 10% that of a man of the same age who had continued to smoke at the same daily level (Doll & Peto, 1976). Thus, if one man starts smoking at age 20 and smokes 10 cigarettes a day, and a second man smokes 20 cigarettes a day between ages 20 and 40 and then stops, by age 60 the latter will have $(20/10) \times 10\% = 20\%$ of the risk of the former. Total cigarette consumption is the same.

These examples illustrate the danger of condensing the different types of information on exposure into a single measure at the start of the analysis. Each facet of exposure should be examined separately, and only combined, if at all, at a later stage in the analysis.

The preliminary analyses associating the factors under study with disease risk will treat each factor separately. For *dichotomous variables*, a simple two-way table relating exposure to disease can be constructed. The frequency of the exposure among the controls together with an estimate of the relative risk, with corresponding confidence intervals, gives a complete summary of the data.

For *qualitative* or, as they are sometimes called, categorical variables, which can take one of a discrete set of values, direct calculation of relative risk is again straightforward. A specific level would be selected as a baseline or reference level, and risks would be calculated for the other levels relative to this baseline. Choice of the baseline level depends on whether the levels are *ordered*, such as parity or birth order, or *unordered*, as in the case of genetic phenotypes. In the latter situation, a good choice of baseline is the level which occurs most frequently. The choice is particularly important when using the estimation procedures which combine information from a series of 2×2 tables, since the estimates of relative risk between pairs of levels can vary depending on which one was selected as baseline (see § 4.5).

For ordered categorical variables, one would often choose either the highest or the lowest level, with infrequently occurring extreme levels perhaps being grouped with the next less extreme.

By choosing an extreme level as baseline, one expects to see a smooth increase (or decrease) away from unity in the relative risk associated with increasing (or decreasing) level of the factor, if the factor plays a role in disease development. In the early stages of an analysis, it is usually bad practice to group the different levels of a categorical variable before one has looked at the relevant risks associated with each level. The risks of overlooking important features of the data more than outweigh the theoretical distortion of subsequent significance levels.

Quantitative variables are those measured on some continuous scale, where the number of possible levels is limited only by the accuracy of the recording system. Variables

of this type can be treated in two ways. They can be converted into ordered categorical variables by division of the scale of measurement, or they can be treated as continuous variables by postulating a specific mathematical relationship between the relative risk and the value of the variables. In preliminary analyses the former approach would usually be employed, since it provides a broad, assumption-free description of the change of risk with the changing level of the factor. The choice of mathematical relationship used in later analysis would then be guided by earlier results.

In deciding on the grouping of continuous variables, the prime objective should be to display the full range of risk associated with the variable, and also to determine the extent to which a dose response can be demonstrated. With these ends in view, the following guidelines are often of value:

1. A pure non-exposed category should be the baseline level if the numbers appear adequate (e.g., more than five to ten individuals in both case and control groups). Thus, to examine the effect of smoking, where consumption might be measured in grams of tobacco smoked per day, a clearer picture of risk is obtained by comparing different smoking categories to non-smokers than by pooling light smokers with non-smokers (Tables 6.6 and 6.8).
2. A simple dichotomy may conceal more information that it reveals. The thirtyfold range in risk for lung cancer between non-smokers and heavy cigarette smokers is greatly obscured if smoking history is dichotomized into, say, one group composed of non-smokers and smokers of less than ten cigarettes a day as opposed to another group of smokers of ten or more cigarettes a day.
3. Use of more than five or six exposure levels will only rarely give added insight to the data. The trends of risk with exposure as defined by a grouping into five levels are usually sufficient. Three levels, in fact, are often adequate, particularly when the data are too few to demonstrate a smooth increase of risk with increasing dose (Cox, 1957; Billewicz, 1965).

Example: Table 3.3 shows relative risks for breast cancer associated with age at first birth among a cohort of 31 000 Icelandic women who had visited a cervical cancer screening programme at least once by 1974 (Tulinius et al., 1978). The lowest risk group, women who gave birth before 20 years of age, is taken as the baseline level. The alternative analysis based on a dichotomy at 25 years is presented for comparison. Even as a preliminary analysis, the greater range of risk, together with the smooth trend, makes the finer categorization of age at first birth considerably more informative.

Table 3.3 Relative risk of breast cancer associated with age at first birth, after adjusting for year of birth, among 31 000 Icelandic women^a

Age at first birth	Relative risk	
<20	1.00	1.00
20-24	1.63	
25-29	2.61	
30-34	2.53	2.05
35+	4.12	
Nulliparous	3.76	

Once the general form of the relationship between exposure level and risk has been ascertained, the change of risk can be modelled in terms of a mathematical relationship. The advantages of using mathematical models for expressing the change in risk over a range of exposure levels are economy in the number of parameters required, and a smoothing of the random fluctuations in the observed data. The advantages, in fact, are those that generally result from using a regression equation to summarize a set of points. This topic is discussed further in § 3.6 and in detail in Chapters 6 and 7.

Further analyses will investigate in a series of stages the combined action of factors of interest. First, we may wish to consider individual factors separately and to examine how the other variables modify their effect. This modification may consist of a general confounding effect, in which the association between the different exposures distorts the underlying disease exposure associations, or of interaction when the exposure risk may be heterogeneous over the different values of the other variables. Second, we may want to examine the joint effect of several exposures simultaneously.

We shall start by consideration of confounding effects.

3.4 Confounding

Confounding is the distortion of a disease/exposure association brought about by the association of other factors with both disease and exposure, the latter associations with the disease being causal. These factors are called confounding factors. One can envisage two simple types of situation. First, we might have a confounding factor that has two levels, in which disease and exposure were distributed as follows:

Confounder	{	Level 1	High risk for disease High prevalence of exposure
		Level 2	Low risk for disease Low prevalence of exposure

As an example, the disease could be lung cancer, the exposure some occupation primarily of blue-collar workers, and the confounder cigarette smoking. At least in the United States, cigarette smoking is considerably more frequent among blue-collar workers than among managers or professional workers.

One can see in this situation that ignoring the confounder will make the association between exposure and disease risk more positive than it would otherwise be. High risk for disease and high prevalence of exposure go together, as do low risk for disease and low prevalence of exposure.

A second type of situation that might arise would be:

Confounder	{	Level 1	High risk for disease Low prevalence of exposure
		Level 2	Low risk for disease High prevalence of exposure

One might take as an example a study relating breast cancer to use of oestrogens for menopausal symptoms, the confounder being age at menopause. Early menopause

decreases breast cancer risk, but leads to greater use of replacement oestrogens (Casa-grande et al., 1976). Here ignoring the confounding variable will make the association between the disease and exposure appear less positive than it should be.

We shall begin our discussion of confounding by a treatment of the statistical concepts, the occasions on which confounding is likely to occur, and to what degree, and the steps that can be taken both in design and analysis to remove the effect of confounding on observed associations. However, confounding cannot be discussed solely in statistical terms. Occasions arise in which the association of one factor with disease appears at least partially to be explained by a second factor (associated both with disease and with the first factor), but where the two factors are essentially measuring the same thing, or where the second factor is a consequence of the first. Under these circumstances, it would be inappropriate to consider the second factor as confounding the association of the first factor with disease. This problem is related to that of overmatching, which we shall consider after we have discussed the statistical aspects of confounding.

Statistical aspects of confounding: dichotomous variables

We shall start by considering two dichotomous variables, one of which we shall regard as the exposure of interest (E), the other a potential confounding variable (C).

Suppose we had obtained, when cross-tabulating disease status against exposure E, the following result based on pooling the data over levels of the confounder (C):

	Exposure E		
	+	-	
Case	a	b	n_1
Control	c	d	n_0
	m_1	m_0	N

As we saw in Chapter 2, the risk ratio associated with exposure to E is well approximated by the odds ratio in the above table.

$$\psi_p = ad/bc$$

where the p subscript means that ψ_p is calculated from the pooled data.

If, now, we consider that the association between E and disease may be partly a reflection of the association of C with both E and disease, then we should be concerned with the association between E and disease for fixed values of C. That is, we shall be interested in the tabulation of disease status against E obtained after stratifying the study population by variable C, as follows:

	Factor C+ Exposure E			Factor C- Exposure E		
	+	-		+	-	
Case	a ₁	b ₁	n ₁₁	a ₂	b ₂	n ₁₂
Control	c ₁	d ₁	n ₀₁	c ₂	d ₂	n ₀₂
	m ₁₁	m ₀₁	N ₁	m ₁₂	m ₀₂	N ₂
	Odds ratio = ψ_1			Odds ratio = ψ_2		

It is clear that the association between E and disease within each of these two 2×2 tables is independent of C since within each table C is the same for all individuals.

We shall assume in this section that $\psi_1 = \psi_2$, i.e., that the association between E and disease is the same in the two strata, and call the common value ψ . In § 3.5, we shall examine situations where this assumption does not hold. Throughout this section, we are considering the odds ratios as population values rather than sample values, so that the equality $\psi_1 = \psi_2$ refers to the underlying population.

The odds ratio ψ represents the association between E and disease after removing the confounding effect of C. Confounding occurs if, and only if, both the following conditions hold:

1. C and E are associated in the control group (which, from the assumption $\psi_1 = \psi_2$, means also in the case group).
2. Factor C is associated with disease after stratification by E.

Factor C is said to confound the association between E and disease status if, and only if, $\psi_p \neq \psi$, that is, if stratifying by C alters the association between E and disease.

These conditions are sometimes loosely expressed by saying that C is related both to exposure and to disease. It should be stressed that the association of C with E must be considered separately for diseased and disease-free persons, and that the association between C and disease must be considered separately among those exposed to E and those not exposed to E.

A distinction is usefully made between confounding effects which create a spurious association and confounding effects which mask a real association. With the former, the crude odds ratio ψ_p will be further from unity than the post-stratification odds ratio ψ . This situation is called *positive* confounding. In the latter situation, the crude odds ratio ψ_p will be closer to unity than the post-stratification odds ratio ψ . This effect is called *negative* confounding. Situations may even arise in which the crude odds ratio is on the opposite side of unity from the post-stratification odds ratio, but they are infrequent.

Confounding, as we have just seen, depends on the association of the confounding variable both with disease and with the exposure, and we can express quantitatively the degree of confounding in terms of the strength of these two associations. In § 2.9, we discussed attributable risk, and the extent to which differences in risk between two populations could be explained by some factor. The situation here is directly analogous; we are considering the degree to which the difference in risk between those exposed to

E and those not exposed to E can be explained by factor C. Equation (2.16) is then directly applicable, and we have:

$$\psi_p = \psi \times \frac{\psi_c p_1 + (1-p_1)}{\psi_c p_2 + (1-p_2)} \quad (3.1)$$

where ψ_c is the odds ratio associating C with disease after stratification by E, p_1 is the proportion of controls among those exposed to E who are also exposed to C, and p_2 is the proportion of controls among those not exposed to E who are exposed to C (see Schlesselman, 1978, for example). When either $\psi_c = 1$ or $p_1 = p_2$, then $\psi_p = \psi$ and there is no confounding effect, giving algebraic expression to the two conditions stated earlier in this section. Expression 3.1 generalizes the result of Cornfield given in § 2.7.

The ratio ψ_p/ψ ($= w$, say) is a measure of the degree of confounding and has been referred to as the *confounding risk ratio* (Miettinen, 1972). Table 3.4 gives the value of the confounding risk ratio for various degrees of association between C and disease, and for different values of p_1 and p_2 . It is of interest to note that the confounding risk ratio is considerably less extreme than the association of either C with disease or C with exposure E. Confounding factors have to be strongly associated with both disease and exposure to generate spurious risk ratios greater than, say, two (see, for example, Bross, 1967). One should stress that the aim of the analysis is not to estimate the con-

Table 3.4 Confounding risk ratios associated with varying relative risk (ψ_c), frequency of occurrence of the confounding variable among controls exposed to E (p_1) and not exposed to E (p_2)

Value of p_2	$\psi_c = 2$			
	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.18	1.36	1.64
0.3	0.85	1	1.15	1.38
0.5	0.75	0.87	1	1.20
0.8	0.61	0.72	0.83	1

Value of p_2	$\psi_c = 5$			
	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.57	2.14	3.00
0.3	0.64	1	1.36	1.91
0.5	0.47	0.73	1	1.40
0.8	0.33	0.52	0.71	1

Value of p_2	$\psi_c = 10$			
	Value of p_1 0.1	0.3	0.5	0.8
0.1	1	1.95	2.80	4.32
0.3	0.51	1	1.49	2.22
0.5	0.35	0.67	1	1.49
0.8	0.23	0.45	0.67	1

founding risk ratio, but to remove the confounding effects. The purpose of Table 3.4 is simply to indicate how large these effects may be.

From (3.1), we see that ψ_p/ψ , the confounding risk ratio, is greater than unity if, and only if, either (a) C is positively associated with both E and with disease ($\psi_c > 1$ and $p_1 > p_2$) or (b) C is negatively associated with both E and with disease ($\psi_c < 1$ and $p_1 < p_2$). Consequently, if the signs of E and C are arranged to make both E and C positively associated with disease, then negative confounding will only occur if the association between E and C is negative. This result is of some value as it provides a mechanism for drawing one's attention to the concealed associations that may result from negative confounding. (Strictly speaking, E and C should be made positively associated with disease after stratification by the other variable; however, in practice, the pre- and post-stratification risk ratios will usually be on the same side of unity).

More general confounding variables

In the previous section we have considered, for a simple dichotomous confounding variable, one of the two major approaches to the treatment of confounding variables, the approach *via* stratification. The extension of this approach to variables taking several levels, or to situations where there are more than two categorical factors under consideration simultaneously, introduces no new conceptual problems. For a polytomous variable which is suspected of being a confounder, one simply stratifies individuals into groups according to the level this variable takes. When several categorical variables are all considered to be potentially confounding, one stratifies simultaneously by them all. For example, if Factor C_1 takes three levels (I, II, III) and Factor C_2 four levels (1, 2, 3, 4), and both are thought to confound the association of Factor E (which we shall take to be dichotomous) with disease, then the data have to be grouped into 12 strata, and the 2×2 tables relating Factor E to disease constructed for each as follows:

Factor C_1		Factor C_2							
		1		2		3		4	
		Exposure E		Exposure E		Exposure E		Exposure E	
		+	-	+	-	+	-	+	-
I Case									
	Control								
II Case									
	Control								
III Case									
	Control								

The confounding effects of C_1 and C_2 have been eliminated, and we can estimate the independent association of E with disease. Methods for constructing summary estimates of the relative risk associated with E, and summary significance tests, are given in the next chapter. Continuous variables can be incorporated into this approach by dividing up the scale of measurement and treating them as ordered categorical variables.

When the confounding variables take more than two levels, the criteria we discussed for assessing when a dichotomous variable might confound an exposure/disease association need to be slightly relaxed. For dichotomous variables, a factor confounds an association if, and only if, it is associated both with disease and exposure. The "only if" part of this criterion holds for all potentially confounding variables, but with polytomous factors we can construct examples in which a factor is related both to disease and to exposure, but does not confound the disease-exposure association (Whittemore, 1978).

There also needs to be some modification of criteria for assessing confounding when more than one confounding variable is present. In the following example, from Fisher and Patil (1974), we have two confounding variables, C_1 and C_2 . Neither one alone confounds the association of E with disease, but the two jointly do confound the association. Stratifying by each of the two possible confounders, in turn, we have:

	Stratification by C_1				Stratification by C_2					
	No stratification		Factor C_1+		Factor C_1-		Factor C_2+		Factor C_2-	
	Exposure E		Exposure E		Exposure E		Exposure E		Exposure E	
	+	-	+	-	+	-	+	-	+	-
Case	12	30	6	15	6	15	6	15	6	15
Control	4	22	2	11	2	11	2	11	2	11
Odds ratio	2.2		2.2		2.2		2.2		2.2	

But, when we stratify by both confounders jointly we have:

	Joint stratification by C_1 and C_2							
	Factor C_1+C_2+		Factor C_1+C_2-		Factor C_1-C_2+		Factor C_1-C_2-	
	Exposure E + -		Exposure E + -		Exposure E + -		Exposure E + -	
Case	1	10	5	5	5	5	1	10
Control	1	10	1	1	1	1	1	10
Odds ratio	1		1		1		1	

The crude association between E and disease, unaffected by stratification by either C_1 or C_2 alone, disappears upon stratification by both confounders simultaneously. One has to distinguish between the individual confounding effects that variables may have, and the joint confounding effects when variables are considered together.

In the latter case, when considering a set of potential confounders, we can see that there is a *joint* effect only if the *joint* distribution of the confounding variables varies with E conditional on disease status, and varies with disease status conditional on E. The corresponding "if" part of this statement need not apply for a set of confounders, in analogous fashion for a single polytomous variable.

Such situations, however, can be regarded as exceptional, and are mentioned mainly for logical completeness. Normal epidemiological practice is to treat any factor related to disease and exposure as a potential confounder, and there would be few occasions on which one would investigate whether the criteria for joint confounding held (Miettinen, 1974).

Degree of stratification

With several confounding variables, or a single confounder with many values, there is the problem of how fine to make the stratification. If the data are divided into an excessive number of cells, information will be lost; but, if the stratification is too coarse then its object will not be achieved and some confounding will remain. Guidelines can be provided by considering the confounding risk ratio resulting from different levels of stratification. Suppose we have a confounding factor C which can take K levels, and after stratification by E the K levels have associated relative risks $r_1 = 1, r_2, \dots, r_K$ (level 1 is baseline). We suppose that these levels occur with frequencies p_{11}, \dots, p_{1K} , respectively, among the controls exposed to the factor of interest E, and frequencies p_{21}, \dots, p_{2K} among controls not exposed to E. As an extension of (3.1) following from (2.16), the confounding risk ratio, which we shall write as w, is the ratio of the odds ratio ψ_p associated with E before stratification by C to the odds ratio ψ after stratification by C, and is given by:

$$w = \frac{\psi_p}{\psi} = \frac{\sum_{k=1}^K p_{1k} r_k}{\sum_{k=1}^K p_{2k} r_k} \quad (3.2)$$

Now suppose the K levels of C are grouped into a smaller number of levels, say J. Since the p_{1k} and p_{2k} are not the same, the risk among one set of pooled levels relative to the risk among the lowest set of pooled levels may differ between those exposed to E and those not exposed to E. Thus, pooling levels of C may have generated an interaction between C and E. This effect, however, will usually be small, and we shall consider the relative risks in those not exposed to E as summarizing the relative risks in the pooled levels of C. The frequency of occurrence of the J pooled levels of C we shall denote by p^*_{1j} among those exposed to E and by p^*_{2j} among those not exposed to E, $j = 1, \dots, J$. The relative risk in the j^{th} pooled level (among those not exposed to E) we shall write as r^*_j , $j = 1, \dots, J$, with $r^*_1 = 1$. The confounding risk ratio for the pooled levels of C, w^* , say, is then given by:

$$w^* = \frac{\sum_{j=1}^J r_j^* p_{1j}^*}{\sum_{j=1}^J r_j^* p_{2j}^*} \quad (3.3)$$

Comparison of w^* with w enables one to assess the extent to which the grouped levels of C remove the full confounding effect. The ratio w/w^* is a measure of the residual confounding effect.

The relationship between w^* and w can be examined more closely. The K levels of C have been classed into J groups. Within each group of levels, if the risk ratios and the ratios (p_{1k}/p_{2k}) are not identical, there will be a residual confounding effect and a corresponding residual confounding risk ratio, which for the j^{th} group we will write as w_j^* ($j = 1, \dots, J$).

We can then express the overall confounding risk ratio w in terms of the w_j^* as:

$$w = \frac{\sum_{j=1}^J w_j^* r_j^* p_{1j}^*}{\sum_{j=1}^J r_j^* p_{2j}^*}$$

which gives, from (3.3)

$$\frac{w}{w^*} = \frac{\sum_{j=1}^J w_j^* r_j^* p_{1j}^*}{\sum_{j=1}^J r_j^* p_{1j}^*} \quad (3.4)$$

The overall measure of residual confounding (w/w^*) is the weighted average over the J groups of the residual confounding risk ratio within each of the grouped levels of C . Computation of the different w_j^* will identify those groups for which finer stratification may be necessary.

As an example, we might consider a study of lung cancer in which interest was focused on an exposure E , other than smoking, and cigarette smoking is to be treated as a confounder. Table 3.5 gives the confounding risk ratio for various possible groupings of cigarette consumption. Data from Doll and Peto (1978) are used for the risk ratios, and the distribution of cigarette consumption among British doctors is used as the distribution among those not exposed to E , i.e., as the values of p_{2k} . The smoking distribution p_{1k} represents a heavy smoking population, an industrially exposed group, for example.

This example has some interesting features. One can see that the grouping 0, 1–9, 10–19, 20–29, 30–40 leaves an inappreciable residual confounding effect and that the grouping 0, 1–19, 20–40 leaves a residual confounding risk ratio of 1.16 at most. A considerable residual confounding effect remains if non-smokers are grouped even with those smoking 1–4 cigarettes a day, underlining the importance of keeping an unexposed group, as stressed in § 3.3. This residual effect makes only a minor contribution to the total confounding effect, as most weight is attached to the heavy smokers (see equation 3.4), but if one is interested in light smokers, this effect is obviously important.

Table 3.5 Residual confounding effects after various degrees of stratification by cigarette consumption

Average daily cigarette consumption	Risk ratio ^a	p _{2i} ^a %	p _{ii} %	Grouping, with residual confounding risk ratio within each group						
				I	II	III	IV	V	VI	VII
0	1	38	10	1.0	1.0	1.39	1.39	1.0	1.57	1.31
1-4	5.6	3	2	0.97	1.05			1.01		
5-9	3.2	8	7			1.01	1.01		1.28	1.16
10-14	9.4	10	9	1.01	1.01			1.41		
15-19	11.3	10	12			1.01	1.01		1.41	1.61
20-24	23.2	15	15	1.01	1.01			1.41		
25-29	24.9	7	15			1.02	1.02		1.41	1.61
30-34	38.2	5	15	1.02	1.02			1.41		
35-40	50.7	4	15			1.02	1.02		1.41	1.61
Confounding risk ratio = 1.93				1.91	1.68			1.86		

^a Data adapted from Doll and Peto (1978)

When stratifying by several confounding variables simultaneously, the joint confounding risk ratio will often be more extreme than either one singly. In fact, if the confounding variables are mutually independent, the joint confounding risk ratio will be the product of the individual confounding risk ratios. If the levels of the different confounding variables are grouped, then the overall residual confounding risk ratio is the product of the individual residual risk ratios. For example, suppose we had two independent confounding variables like that in Table 3.5. The joint confounding risk ratio would be $(1.93)^2 = 3.72$. If both variables are grouped as in column II, the individual residual confounding risk ratios are $1.93/1.68 = 1.15$, whereas the joint residual confounding risk ratio is $(1.15)^2 = 1.32$. Thus for the same level of stratification, the residual confounding effect tends to increase with the number of confounding variables. An increasing penalty is paid if one yields to the temptation to coarsen the stratification as the confounding variables increase in number. The control of confounding by stratification clearly runs into trouble as the number of confounding variables increases, unless one has very large samples. What is required is a method, after a relatively fine stratification of each variable, of combining different strata into roughly homogeneous groups. Various *ad hoc* methods have been proposed, such as the sweep and smear technique (Bunker et al., 1969) or the confounder score index (Miettinen, 1976), but these methods can give incorrect answers (Scott, 1978; Pike, Anderson & Day, 1979), and the unified approach *via* logistic regression is recommended (see Chapters 6 and 7).

Effect of study design on confounding effects

An extreme but not unusual example of positive confounding would be data such as the following:

	Factor C+ Exposure E		Factor C- Exposure E		Pooled levels of C Exposure E	
	+	-	+	-	+	-
Case	90	10	1	9	91	19
Control	9	1	10	90	19	91
Odds ratio	1		1		22.9	

From these tables it can be seen that the cell entries when C is positive differ markedly from the cell entries when C is negative. This lack of balance has two consequences. First, the unequal ratio of cases to controls and the unequal proportion of those exposed to E, in the two post-stratification tables, lead to strong positive confounding. Second, as a reflection of the confounding the minimum cell entries in the two tables obtained after stratification are both much smaller (both equal to one, in fact) than half the minimum cell entry in the pooled table (equal to 19). Thus one can expect the estimate of the odds ratio to be considerably less precise than an estimate obtained from more balanced tables. Both effects can be mitigated by equalizing the ratio of cases to controls in those exposed and those not exposed to a confounder C, in which case we say the design is balanced for Factor C. The results could be represented as follows:

	Factor C+ Exposure E			Factor C- Exposure E			Pooled levels of C Exposure E		
	+	-		+	-		+	-	
Case	a_1	b_1	n_1	a_2	b_2	n_2	$a_1 + a_2$	$b_1 + b_2$	$n_1 + n_2$
Control	c_1	d_1	mn_1	c_2	d_2	mn_2	$c_1 + c_2$	$d_1 + d_2$	$m(n_1 + n_2)$
Odds ratio	ψ			ψ			ψ_p		

Balancing or even equalizing cases and controls in each stratum does not eliminate confounding, as the following example illustrates:

	Factor C+ Exposure E		Factor C- Exposure E		Pooled levels of C Exposure E	
	+	-	+	-	+	-
Case	50	50	90	10	140	60
Control	10	90	50	50	60	140
Odds ratio	9		9		5.44	

However, using expression (3.1) and the balance in the design one can show that the odds ratio in the pooled table, ψ_p , lies between unity and ψ , that is we have:

$$1 < \psi_p < \psi, \text{ if } \psi > 1$$

or

$$1 > \psi_p > \psi, \text{ if } \psi < 1.$$

Thus, in contrast to the unbalanced situation, where the confounding effects can be either positive or negative and where the pooled odds ratio can be the opposite side of unity from the within-stratum odds ratio, *with a balanced design the expected pooled odds ratio is both on the same side of and closer to unity than the expected within-stratum odds ratio*. An unstratified analysis will bias the odds ratio towards unity, unless the confounding and exposure variables are (conditional on disease status) independent, but not change the side of unity on which the odds ratio lies.

Obviously the odds ratio of C with disease bears no relation to the true association. *If a factor has been balanced, the data so generated give no information on the association of that factor with disease*. Interaction can still be estimated, however, as is discussed in § 3.5.

Balance, as described above, where the control series is chosen to ensure equal frequency of cases and controls in different strata, is sometimes referred to as frequency matching. On other occasions, where for each case a set of controls is chosen to have the same, or nearly the same, values of prescribed covariates, we speak of individual matching. In later chapters matching refers specifically to individual matching.

Incorporation of matching factors in the analysis

The purpose of matching, as we have just seen, is to control confounding and increase the information per observation in the post-stratification analysis. Most studies, and certainly those of cancer, would match for age and sex, since both could confound the effect of most other factors. A large number of studies match on additional variables, often to the point where each case may be associated with a set of controls in an individual stratum. One purpose of this matching, as we have mentioned, is to improve the precision of the estimates of the relevant relative risks obtained from a stratified analysis. Some matching factors, such as place of residence or membership of a sibship, represent a complex of factors. Then, the purpose of the matching is to eliminate the confounding effect of a range of only vaguely specified variables, since the matching provides a stratification by these variables which would otherwise be difficult to perform because of their indeterminate nature. In these circumstances, matching can be an important way of eliminating bias in the risk estimate. The result given for a dichotomous matching variable can be extended without difficulty to any complexity of matching variables. The expected odds ratio resulting from an analysis incorporating the matching is always more extreme than the expected odds ratio obtained ignoring the matching (Seigel & Greenhouse, 1973).

Now, the purpose of matching implies that the matching factors must *a priori* be considered as ones for which stratification would be necessary, that is, as confounding variables. It would follow *that variables which have been used for matching in the design should be incorporated in the analysis as confounding variables*. Until recently,

there were limitations on the type of analysis that could be done which fully incorporated the matching. However, the analytical methods now available do not suffer from these limitations.

The extent to which the analysis should incorporate the matching variables will depend on how the variables are used for matching. If matching is performed only on age and sex then a stratified analysis rather than one which retains individual matching may be more appropriate. Individual matching in the analysis is only necessary if matching in the design was genuinely at the individual level. However, preservation of individual matching, even if artificial, can sometimes have computational advantages and often means little loss of information (see § 7.6).

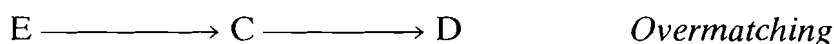
Overmatching

It might be inferred from our discussion that the post-stratification analysis is always the one of interest, that if we can find a variable which appears to alter the association between disease and the exposure then we should treat that variable as a confounder, but this approach ignores the biological meaning of the variables in question and their position in the sequence of events which leads to disease.

A diagrammatic representation of (positive) confounding might be as follows:



In many situations, however, such a figure does not correspond to the true state of affairs. Two such situations merit particular attention. The first is when an apparent confounding variable in fact results from the exposure it appears to confound. We could represent this occurrence diagrammatically as



where C is part of the overall pathway



Chronic cough, smoking, and lung cancer can be cited as an example. One would expect the pattern of cigarette smoking among those with chronic cough to be closer to the smoking pattern of lung cancer cases than to that of the general population. The result of stratifying by the presence of chronic cough before diagnosis of lung cancer might almost eliminate the lung cancer-smoking association. The real association between smoking and lung cancer is obscured by the intervention of an intermediate stage in the disease process. A similar example is given by cancer of the endometrium, use of oestrogens and uterine bleeding. If use of oestrogens by postmenopausal women induces uterine bleeding, itself associated with endometrial cancer, then one might expect stratification by a previous history of uterine bleeding to reduce the association between endometrial cancer and oestrogens. If history of chronic cough or a history of uterine bleeding were used as stratifying factors in the respective analyses, or as a matching factor in the design, then one would call the resulting reductions in the

strength of the disease/exposure association examples of overmatching. In both these examples, the overmatching consists of using as a confounding factor a variable whose presence is caused by the exposure.

A second way in which overmatching may occur is when both the exposure and the confounder represent the same underlying cause of the disease. We might represent such a situation as:



C and E now represent different aspects of the same composite factor causally related to disease. For example, C and E might both be aspects of dietary fibre, or alternative measures of socioeconomic status. From the diagram it is clear that both should have equal status as associates of disease. One might, somewhat arbitrarily, decide to take one of the two, or even attempt to form a composite variable using regression methods. It would clearly be inappropriate to consider one as confounding the effect of the other, or to consider the association of one with disease after stratification by the other.

In both the above situations, overmatching will lead to biased estimates of the relative risk of interest.

A third way in which overmatching may occur is through excessive stratification. The standard errors of post-stratification estimates of relative risk tend to be larger than the standard errors of pre-stratification estimates (see § 7.0). Stratification by factors which are not genuine confounding variables will therefore increase the variability of the estimates without eliminating any bias, and can be regarded as a type of overmatching. It is commonly seen when data are stratified by a variable known to be associated with exposure but not in itself independently related to disease. It does not give rise to bias. If one recalls the section of Chapter 1 relating to overmatching in the design of a study, one can see close parallels between the different manifestations of overmatching in the design and the analysis of a study.

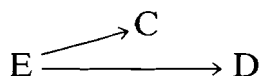
We may represent diagrammatically the situation in which a variable is related to exposure but not to disease as:



C is not a genuine confounding variable. Simply by chance, however, substantial confounding may appear to occur, as the result of random sampling. This eventuality will arise more often the more strongly E and C are associated. It may be difficult to decide whether such an event has occurred, and this will normally require consideration of how C and disease could, logically or biologically, be related. If several studies have been performed, such confounding may appear as an inconsistency in the results, with different factors appearing to have confounding effects in different studies. Good evidence may be available from previous studies that C is not causally related to disease, in which case it should not be incorporated as a confounder. If, nevertheless, it appears

to have a strong confounding effect, the design of the study should be carefully examined to ensure that it is not acting as a surrogate for some other potential confounder, and in particular that it is not acting through selection bias (see § 3.7).

Note that the situation



may lead to genuine confounding when the variables are measured with error. Consider as an example:

	Factor C+ Exposure E		Factor C- Exposure E		Pooled levels of C Exposure E	
	+	-	+	-	+	-
Case	90	1	10	9	100	10
Control	9	10	1	90	10	100
Odds ratio	100		100		100	

Suppose now that E is misclassified 10% of the time, yielding a variable which we shall denote by E*. The tables become, approximately:

	Factor C+ Exposure E*		Factor C- Exposure E*		Pooled levels of C Exposure E*	
	+	-	+	-	+	-
Case	81	10	10	9	91	19
Control	9	10	10	81	19	91
Odds ratio	9		9		22.9	

The post-stratification odds ratio relating E* to disease is much less than that relating E to disease but in addition a confounding effect has arisen, with a confounding risk ratio of $22.9/9 = 2.54$. The odds ratio relating C to disease, after stratification by E*, is 9 rather than 1. The reason is clear: both E* and C are correlates of E, and both are related to disease only through E. Only if E is exactly known does knowledge of C contribute nothing extra to assessment of disease risk.

It is clear from our discussion of confounding that it is not an issue which can be settled on statistical grounds. One has to consider the nature of the variables concerned, and of their relationships with each other and with disease.

Variables to be included as confounding variables

We have considered the conditions under which an observed association may be the result of a confounding effect, and when overmatching might occur, and have discussed the criteria for deciding which factors to incorporate in the analysis as confounding

factors when confronting the data from a particular study. Normally there will be two basic aims: first, to remove from the disease/exposure associations of interest all the confounding effects present in the study data set, whether positive or negative; second, to ensure that genuine associations are not reduced by overmatching.

To satisfy the first aim, questions of statistical significance are irrelevant. Given that a confounding factor has to be associated both with disease and with exposure, one might contemplate testing whether both associations are significant in the available data. If their association were not significant, then one might discard the factor as a potential confounder. *This approach is incorrect* (Dales & Ury, 1978), and it can lead to substantial confounding effects remaining in the association, as the following example shows:

Stratification by potential confounder

	Factor C+ Exposure E		Factor C- Exposure E		Pooled levels of C Exposure E	
	+	-	+	-	+	-
Case	80	40	5	5	85	45
Control	8	4	40	40	48	44
Odds ratio	1		1		1.73 ($\chi^2 = 3.91$)	

The association between E and C, after cross-classification by disease status, does not achieve significance at the 5% level: $\chi^2 = 1.63$ using the Mantel-Haenszel χ^2 given by equation (4.23). Thus, in the data C and E are not significantly associated, but an appreciable and statistically significant (at least in the formal sense) association between E and disease exists before stratification by C, which vanishes upon stratification.

With this example in mind, and recalling the initial discussion of overmatching, we can propose three criteria for treating a variable as a confounding variable in the analysis.

1. If a variable C is known from other studies to be related to disease, and if this association is not subsidiary to a possible exposure/disease association, then C should be treated as a confounding variable. The significance of the association between C and disease in the data at hand is of no relevance. Irrespective of the association between E and C in the general population, if there is an association between E and C in the study sample then part of the association between E and disease in the study sample will be a reflection of the causal association between C and disease. The contribution of C to the E-disease association must be eliminated before proceeding to further considerations of a possible causal role for E in disease development. Age and sex will almost always be confounding variables, and should be treated as such.
2. If a variable C is related to disease, but this association is subsidiary to the association between E and disease, by which we mean that either C is caused by E or forms a part of the chain of events by which disease develops from E, then C should not be considered as a confounder of the disease/exposure association.

3. If a factor is thought important enough to be incorporated in the design of the study as a matching or balancing factor, then it should be treated as a confounding variable in the analysis.

In the situation when E and C are known to be related, and if in the data C is also related to disease, then there will be an apparent confounding effect. In this situation, unlike the previous one, it is less clear what the interpretation should be in terms of causality. Incorporating C as a confounding variable implies that one is giving the C-disease association precedence over the E-disease association, which one would not always want to do, as for example when C and E are different measures of the same composite factor. The possibility must be considered that selection bias has operated with respect to C in the choice of either cases or controls or in the manner of acquiring information. Control of this bias may be possible by treating it as if it were a confounding effect. This is discussed in § 3.7.

3.5 Interaction and effect modification

In our discussion of the joint effect of different factors, and specifically in the context of confounding, we have assumed that the odds ratio associating one factor to disease is unaltered by variation in the value of other factors. This simple assumption can only be an approximation, although as we saw in Chapter 2, on many occasions the approximation is fairly close. On other occasions, appreciable variations in the odds ratio were noted, and these variations themselves were of biological importance.

If the odds ratios associating factor A and disease vary with the level of a second factor B, then it is common epidemiological parlance to describe B as an effect modifier. The term is not a particularly happy one, however. A departure from a multiplicative model might arise, for example, if two factors operated in the same way at the cellular level and their joint effect were additive, which would make little sense biologically to describe as effect modification. We prefer to use the term 'interaction', in keeping with usual statistical terminology.

The main reasons for studying interactions are first because they may modify the definition of high risk groups, and second because they may provide insight into disease mechanisms. Interaction implies that in certain subgroups the relative risk associated with exposure is higher than in the rest of the population. Both the specificity of risk for these subgroups, and the fact that the level of exposure-associated risk will be higher than the general risk in the population would tend to increase one's belief in the causal nature of the association, as was discussed earlier in the chapter. The aim should not be to eliminate interactions by suitable transformations, but rather to understand their nature; this point is well made by Rothman (1974).

One should note that using a variable as a matching factor in the design, so that its individual effect on risk cannot be studied, does not alter the interactive effects that the factor may have with other exposures. A simple example will illustrate the point.

	Factor B+ Factor A			Factor B- Factor A		
	+	-		+	-	
Case	50	50	100	25	25	50
Control	5	5	10	20	80	100
Odds ratio	1			4		

Suppose now that the ratio of controls to cases were the same for each level of B. The results would then be, for example:

	Factor B+ Factor A			Factor B- Factor A		
	+	-		+	-	
Case	50	50	100	25	25	50
Control	50	50	100	10	40	50
Odds ratio	1			4		

Thus, the two odds ratios relating factor A to disease for the two levels of B are unaltered, but the odds ratio relating B to disease is greatly modified. Matching does not alter interaction effects between variables used for matching and those not so used.

Analysis of interaction effects

The first step is to investigate if appreciable interactions are present. In the simplest situation, one may just wish to test whether the relative risks in two groups, defined perhaps by age or some other dichotomized variable, are the same; the type of test proposed in Chapter 4 would be appropriate. In more complex situations, two approaches are possible. First, the observed distribution of the exposure variables among the cases and controls can be compared with the distribution under the multiplicative model (see § 2.6). Patterns in the departures of observed from expected may indicate the superiority of an alternative model for the joint action. With two or three variables which can be stratified into a few categories each, the presentation is simple, as shown in Table 3.6 for data relating oral cancer risk to use of tobacco and alcohol (from Wynder, Bross & Feldman, 1957; see also Rothman, 1976). The expected values, obtained using unconditional maximum likelihood methods (Chapter 6, particularly § 6.6) enable one to examine the adequacy of the overall fit of the multiplicative model. In Table 3.6, the fit is good. A feature of Table 3.6 is that apparently substantial differences in the odds ratios can arise from fairly small differences between observed and expected numbers of cases and controls in a cell. The cell corresponding to <1 units of alcohol/day and 34+ cigarettes/day is an example.

Table 3.6 Risk for oral cancer associated with alcohol and tobacco^a

A. Observed and expected number of cases in each smoking and drinking category, with the observed number of controls

	Alcohol (average consumption in units/day)			Tobacco (cigarettes/day)								
	≤15			16–20			21–34			34+		
	Obs. Cont.	Obs. Cases	Exp. Cases	Obs. Cont.	Obs. Cases	Exp. Cases	Obs. Cont.	Obs. Cases	Exp. Cases	Obs. Cont.	Obs. Cases	Exp. Cases
<1	31	16	18.84	19	25	22.54	13	12	13.67	3	13	10.96
1–2	8	7	6.07	18	19	19.10	5	5	5.51	5	10	10.32
3–6	8	20	18.47	16	40	42.14	5	24	22.54	4	19	19.84
6+	2	10	9.62	5	30	30.22	4	35	34.28	4	40	40.88

B. Observed and expected relative risks in each smoking and drinking category

	Alcohol (average consumption in units/day)		Tobacco (cigarettes/day)					
	≤15		16–20		21–34		34+	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
<1	1.00	1.00	2.55	1.57	1.79	1.80	8.41	3.24
1–2	1.70	1.00	2.05	1.57	1.94	1.80	3.88	3.24
3–6	4.85	2.85	4.85	4.47	9.31	5.10	9.21	9.23
6+	9.70	6.03	11.64	9.47	16.98	10.85	19.37	19.53

^a From Wynder, Bross and Feldman (1957)

If significant or appreciable interaction terms are present, one would then attempt to understand the nature of their effect. First, examination of the discrepancies between the observed numbers and those expected under the no-interaction model may indicate that the interaction corresponds to unsystematic but excessive variation. It is a common feature of epidemiological data to show variation slightly higher than that theoretically expected on the basis of random sampling considerations. Since the excess variation probably arises from minor unpredictable irregularities in data collection, one would suspect that it is due mainly to chance, augmented perhaps by variation of extraneous factors.

A second interpretation of the departures from a multiplicative model may be that the variables interact in a different way. An obvious alternative would be to try to fit an additive model for the relative risks. The data on occupational exposure, cigarette smoking and bladder cancer from Boston (Cole, 1973) suggest a better fit for an additive model, as do data on use of oestrogens at the menopause, obesity and risk for endometrial cancer (Mack et al., 1976).

A third interpretation is that specific groups, as defined by the interactive factors, are at higher risk due either to greater susceptibility or to greater exposure. An example of greater susceptibility with age at exposure is provided by the variation in risk for breast cancer due to irradiation (Boice & Stone, 1978). An example of differences in exposure is given by the risk for cancer of the lung and nasal sinuses among nickel refinery workers in South Wales (Doll, Mathews & Morgan, 1977). The risk appears

confined to those first employed before 1930. Changes in the operation of the refinery at that time could, quite plausibly, have removed the carcinogenic agents, and the change in risk assists in identifying what those agents may have been.

3.6 Modelling risk

The use of stratification and cross-tabulation to investigate the joint effect on risk of two variables, in terms of how the two factors mutually confound each other and interact, is reasonably straightforward. However, even with two variables, as the number of values each variable can take increases, the control of confounding by means of stratification can lead to substantial losses of information, and tests for interaction will lack power. As the complexity of the problem increases, the approach *via* stratification becomes not only unwieldy but increasingly wasteful of information. The different effects associated with different levels of a variable will not normally be unrelated, and can be expected to change smoothly. For example, for a quantitative variable, risk will usually vary in a manner which can be described by a simple family of curves. It would be rare to need more than second degree terms, after perhaps some initial transformation of the scale.

Similarly, interactive effects between several variables will not normally vary in a structureless way, and general experience has been that most situations are well described by some simple structure of the interactions. These considerations lead to the use of regression methods, in which the risk associated with each variable is expressed as some explicit function, and interaction effects are described in terms of the specific parameters of interest. Chapters 6 and 7 are devoted to the development of these methods, which will not be further discussed here except to outline briefly their advantages. These can be summarized as follows:

1. One can study the joint effect of several exposures simultaneously. The stratification approach we considered earlier places the emphasis on one specific exposure. Study of the combined risk associated with several exposures is an important complement of the single exposure analysis.
2. When the number of levels of the confounding variables increases, one can remove their effect as fully as by fine stratification but with less loss of information.
3. One can test for specific interaction effects of interest with the considerable increase in power this provides. One also obtains a parametric description of the interaction.
4. The risk associated with different levels of a quantitative variable can be expressed in simple and descriptive terms.

In studies where controls are individually matched to cases, these advantages are accentuated, as Chapters 5 and 7 make apparent. But regression methods should not replace analyses based on cross-tabulation, rather they should complement and extend them, as we illustrate in Chapter 6.

3.7 Comparisons between more than two groups

So far, we have considered methods of analysis appropriate for comparisons between one case group and one control group. Situations occur, however, when comparisons among more than two groups are required. One may want to test whether the relative

risk for some factor is the same over different subcategories of disease, for example, different histological types of lung cancer or different subsites of the oesophagus or the large bowel. Or one may want to test whether the results obtained using different control groups can be taken as equivalent, or whether observed differences are easily explained as chance phenomena. The approach most commonly taken is an informal one, in which one calls attention to appreciable differences in risk when comparisons are made between different pairs of groups, but one does not attempt a formal test of significance.

The methods presented in Chapters 6 and 7 can be extended to the comparison of more than two groups. In particular, if the study design incorporated individual matching, then a dummy variable, indicating disease subcategory could be introduced and interaction examined between this dummy variable and the exposure of interest. However, investigating heterogeneity of disease subcategory or of type of control group by introducing interaction terms is only appropriate if, for each stratum, every individual belongs to one disease subcategory or one control group. If within a stratum more than two groups are represented, then the underlying probability structure needs extension. One cannot simply write

$$\text{pr}(\text{control}) = 1 - \text{pr}(\text{case}).$$

One has to generalize this expression to

$$\sum_i \text{pr}(\text{case, disease category } i) + \sum_j \text{pr}(\text{control, type } j) = 1$$

Mantel (1966) and Prentice and Breslow (1978) have indicated how a generalized logistic function, appropriate to this situation, can be formed and how the various estimation and hypothesis-testing procedures can be derived. The only difficulty in practical use is that the number of parameters can become unwieldy.

One would anticipate that the generalized logistic model will be used more in the future than it has been in the past, at least in the context of epidemiological studies. But, in this monograph we will not consider its development any further, since computer programmes are not readily available to put the techniques in operation, and no important matter of principle is involved. Extensions of the regression models of Chapters 6 and 7 are conceptually simple, and with some labour could be made operational.

3.8 Considerations affecting interpretation of the analysis

The interpretation of a study will depend not only on the numerical results of various analyses, but also on more general considerations of how the study was conducted, the nature of the factors under investigation and the consistency with other studies done in the same field. We conclude this chapter by a brief discussion of some of these aspects.

Bias

Bias is a property of the design of a study and was discussed in Chapter 1. Various features can be incorporated in the design which permit at least partial assessment of

the extent of possible bias; these features are selection of different types of control groups, different series of cases, or different ways of obtaining similar information.

One can include in the study, for example, cases of cancers at different sites, to demonstrate whether the observed effects are specific for the site of interest. An example of this procedure is given by Cook-Mozaffari et al. (1979) in a study of oesophageal cancer in Iran. Frequent practice is to use both a hospital and a population series of controls. If recall and selection bias are present then they should be different in the two groups, and the divergence between the results should indicate the extent of the biasing effect.

The effects of design bias can sometimes be controlled in the analysis in the same way as one controls for confounding. Thus suppose that, whether by selection bias or differential recall, the relative representation of some factor in the case and control groups is different from the relative representation in the study population, and that this factor is related to the exposure of interest. Then, even if the factor does not confound the exposure/disease association in the underlying population, an apparent confounding effect will be seen in the data at hand which can be controlled in the same way as other confounding effects. Care must be taken to ensure that overmatching does not result. Stratification or adjustment by a factor leads to an increase in the variance of the estimates of parameters of interest and routine adjustment by factors for which there is no reason to suspect bias will lead to a loss of information.

More generally, the effects of bias will often be the creation of apparent confounding effects, but little information may be available on the specific variable involved. As in § 3.4, one can assess the strength of association that must exist between these hidden confounding variables and both disease risk and the exposure of interest for the observed association to be due to bias. Biases have to be strong to generate relative risks greater than twofold.

Missing data

Sufficient control should normally be exercised over the conduct of a study to ensure that few, if any, of the data that one intended to collect are missing. Sometimes, however, source documents such as hospital records may be lost or otherwise missing, or respondents may be unable to supply the information required or be unwilling to answer certain questions. The essential point when faced with missing data is to be aware that the occurrence is usually not a random event. The probability that data are missing will be associated with the exposures one is studying, or with disease status or both. If information on some variable is unavailable for any more than a small percentage of individuals, then inferences about that variable will be of doubtful value.

Common practice is to eliminate from analyses including a certain variable all individuals for whom information on that variable is missing. In a matched pair design, the individual matched to an eliminated individual will also be eliminated. Since the individuals eliminated will often form a selected group, their elimination can lead to a biased estimate of relative risk. An attempt can be made to estimate the degree of bias involved. One approach is to replace missing values by the two extreme values, thus bracketing the true results. An alternative, for categorical variables, is to create

an additional category for missing data. The relative risk for this category will give an indication of the degree of bias.

If one can be sure that no bias relates to the absence of data, for example, the accidental loss of biological samples, then one might consider using some of the powerful techniques that have recently been developed for data missing at random (Demster, Laird & Rubin, 1977). These techniques have the advantage that an individual's entire record is not discarded just because a single datum is missing.

Errors of classification

Errors in classifying disease status and in measuring the exposure variables may both be appreciable. For the former, it has long been realized that a certain proportion of the controls may be at an early stage of disease and should have been diagnosed as cases. Under most circumstances, one might expect this proportion and hence the effect to be small, but when the disease under study is particularly common, the effect may become appreciable. The consequence is the same as when misdiagnosed cases are included in the disease group. Both errors lead to underestimates of the relative risk.

Errors of measurement in the exposure variables have been considered by several authors (Bross, 1954; Newell, 1963; Goldberg, 1975; Barron, 1977). Here the effect is also to reduce the apparent risk, unless the errors are linked in some unusual manner to confounding variables. For a dichotomous exposure variable, with a probability Φ of misclassifying an exposed case or control and a probability ϑ of misclassifying a non-exposed person, we have:

$$\text{True odds ratio } \psi = p_1(1-p_0)/p_0(1-p_1)$$

but

$$\text{Observed odds ratio} = \frac{(p_1 + \vartheta/d) \{(1-p_0) + \Phi/d\}}{(p_0 + \vartheta/d) \{(1-p_1) + \Phi/d\}}$$

where p_1 is the true probability of exposure among cases, p_0 the exposure probability for controls, and $d = 1 - \Phi - \vartheta$.

For example, if $p_0 = 0.3$ and $p_1 = 0.1$, with the true relative risk $\psi = 3.9$, then misclassification rates of 10% ($\Phi = \vartheta = 0.1$) will reduce the relative risk to 2.4, and misclassification of 20% reduces the relative risk to 1.7. It should be remembered that sometimes one is not measuring precisely the factor of interest, and that for this reason assessments of the importance of a factor may be too modest. Dietary items would be clear examples; current epidemiological methods are certainly inadequate for estimating retrospectively the intake of dietary fibre or animal fat. Demonstration of an effect for either of these items in a case-control study would be virtually all one could expect, and estimates of relative risk, or attributable risk, are likely to be serious understatements of the real effect. As another example, weak associations (i.e., relative risks of 1.5 or 2) between HLA antigens and disease, particularly locus A or B antigens, are often interpreted as indicating the existence of genes at other loci of the HLA region which are strongly associated with disease.

The problem of multiple comparisons

The results of studies are often presented as if the only variables included in the study were those which *a posteriori* showed association with disease. One has to distinguish between factors which are clearly part of the main hypothesis motivating the study, and those which were included in the study for less obvious reasons. Disease associations demonstrated for the former can be interpreted without consideration of multiple comparisons, but for the latter one or two factors out of a hundred in the study are expected to be significant at the 1% level (this issue is well discussed in the paper by Mantel and Haenszel, 1959). The area where this consideration is most explicitly acknowledged is in the study of genetic polymorphisms and disease, perhaps because non-confirmatory repeat studies are relatively easy to perform. In studies of HLA antigens, for example, it is now required practice to correct nominal significance levels obtained from the 2×2 tables for each antigen by the number of antigens tested (i.e., comparisons). This correction may in fact be over-conservative, especially if interest is mainly directed at the more common antigens, but correction procedures based on, say, a full Bayesian analysis would probably be too complex. A test of association not requiring correction for the number of antigens, based on the multivariate estimates of gene frequencies, has been proposed in this context (Rogentine et al., 1972).

In other areas, when a study has investigated many factors, a few of which achieve nominal significance in the analysis, the interpretation must be cautious, and further studies would be needed to confirm the association. A good example is given by the first report of an association between use of reserpine and risk for breast cancer (Boston Collaborative Drug Surveillance Program, 1974). This study emanated from the Boston Drug Surveillance Program, and the reported association was one out of several hundred possible comparisons (i.e., perhaps 20 tumour sites and 10 or 15 different drugs). As the result on its own was thus uninterpretable, publication was delayed until two further studies had been performed (Armstrong, Stevens & Doll, 1974; Heinonen et al., 1974).

It must always be recognized that no study can be regarded in isolation; the results of each must be viewed in the light of all other relevant information.

REFERENCES

- Aksoy, M., Erdem, S. & Dinçol, G. (1974) Leukemia in shoe workers exposed chronically to benzene. *Blood*, 44, 837–841
- Anderson, D.E. (1974) Genetic study of breast cancer: identification of a high risk group. *Cancer*, 34, 1090–1097
- Armstrong, B., Stevens, N. & Doll, R. (1974) Retrospective study of the association between use of Rauwolfia derivatives and breast cancer in English women. *Lancet*, ii, 672–675
- Barron, B.A. (1977) The effects of misclassification on the estimation of relative risk. *Biometrics*, 33, 414–418
- Billewicz, W.Z. (1965) The efficiency of matched samples: an empirical investigation. *Biometrics*, 21, 623–643

- Boice, J.D. & Monson, R. (1977) Breast cancer in women after repeated fluoroscopic examinations of the chest. *J. natl Cancer Inst.*, 59, 823–832
- Boice, J.D. & Stone, B.J. (1978) *Interaction between radiation and other breast cancer risk factors*. In: *Late Biological Effects of Ionizing Radiation*, Vol. I, Vienna, International Atomic Energy Agency, pp. 231–249
- Boston Collaborative Drug Surveillance Program (1974) Reserpine and breast cancer. *Lancet*, ii, 669–671
- Bross, I. (1954) Misclassification in 2×2 tables. *Biometrics*, 10, 478–486
- Bross I.J. (1967) Pertinency of an extraneous variable. *J. chron. Dis.*, 20, 487–497
- Bross, I. & Tidings, J. (1973) Another look at coffee drinking and cancer of the urinary bladder. *Prev. Med.*, 2, 445–451
- Bunker, J.P., Forrest, W.H., Mosteller, F. & Vandam, L.D., eds (1969) *The National Halothane Study*, Washington DC, US Government Printing Office
- Casagrande, J., Gerkins, V., Henderson, B.E., Mack, T. & Pike, M.C. (1976) Exogenous estrogens and breast cancer in women with natural menopause. *J. natl Cancer Inst.*, 56, 839–841
- Cole, P. (1971) Coffee drinking and cancer of the lower urinary tract. *Lancet*, i, 1335–1337
- Cole, P. (1973) *A population-based study of bladder cancer*. In: Doll, R. & Vodopija, I., eds, *Host Environment Interactions in the Etiology of Cancer in Man*, Lyon, International Agency for Research on Cancer (*IARC Scientific Publications No. 7*), pp. 83–87
- Cook-Mozaffari, P.J., Azordegan, F., Day, N.E., Ressicaud, A., Sabai, C. & Aramesh, B. (1979) Oesophageal cancer studies in the Caspian littoral of Iran: results of a case-control study. *Br. J. Cancer*, 39, 293–309
- Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. natl Cancer Inst.*, 22, 173–203
- Cox, D.R. (1957) A note on grouping. *J. Am. stat. Assoc.*, 52, 543–547
- Dales, L.G. & Ury, H.K. (1978) An improper use of statistical significance testing in studying covariables. *Int. J. Epidemiol.*, 7, 373–377
- Demster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. stat. Soc. B.*, 39, 1–38
- Doll, R. (1969) *Preface*. In: Kreyberg, L., ed., *Aetiology of Lung Cancer*, Oslo, Universitetsforlaget, pp. 9–10
- Doll, R. (1971) The age distribution of cancer. Implication for models of carcinogenesis. *J. R. stat. Soc. A*, 134, 133–155
- Doll, R. & Peto, R. (1976) Mortality in relation to smoking: 20 years' observation on male British doctors. *Br. med. J.*, ii, 1525–1536
- Doll, R. & Peto, R. (1978) Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and life long non-smokers. *J. Epidemiol. Community Health*, 32, 303–313
- Doll, R., Mathews, J.D. & Morgan, L. G. (1977) Cancer of the lung and nasal sinuses in nickel workers: a reassessment of the period of risk. *Br. J. ind. Med.*, 34, 102–105
- Fisher, L. & Patil, K. (1974) Matching and unrelatedness. *Am. J. Epidemiol.*, 100, 347–349

- Fraumeni, J.F., Scotto, J. & Dunham, L.J. (1971) Coffee drinking and bladder cancer. *Lancet*, *ii*, 1204
- Goldberg, J. (1975) The effect of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J. Am. stat. Assoc.*, *70*, 561–567
- Heinonen, O.P., Shapiro, S., Tuominen, L. & Turunen, M.I. (1974) Reserpine use in relation to breast cancer. *Lancet*, *ii*, 675–677
- Herbst, A.L., Ulfelder, H. & Poskanzer, D.C. (1971) Adenocarcinoma of the vagina. *New Engl. J. Med.*, *284*, 878–881
- Hill, A.B. (1965) The environment and health: association or causation. *Proc. R. Soc. Med.*, *58*, 295–300
- Hoover, R. & Fraumeni, J.F. (1973) Risk of cancer in renal transplant recipients. *Lancet*, *ii*, 55–57
- Infante, P.F., Rivisky, R.A. & Wagoner, J.F. (1977) Leukaemia in benzene workers. *Lancet*, *ii*, 76–78
- Jick, H., Watkins, R.M., Hunter, J.R., Dinan, B.J., Madsen, S., Rothman, K.J. & Walker, A.M. (1979) Replacement estrogens and endometrial cancer. *New Engl. J. Med.*, *300*, 218–222
- Mack, T., Pike, M.C., Henderson, B.E., Pfeffer, R.I., Gerkins, V.R., Arthur, M. & Brown, S.E. (1976) Estrogens and endometrial cancer in a retirement community. *New Engl. J. Med.*, *294*, 1262–1267
- MacLennan, R., Da Costa, J., Day, N.E., Law, C.H., Ng, Y.K. & Shanmugaratnam, K. (1977) Risk factors for lung cancer in Singapore Chinese, a population with high female incidence rates. *Int. J. Cancer*, *20*, 854–860
- Mantel, N. (1966) Models for complex contingency tables and polychotomous dosage response curves. *Biometrics*, *22*, 83–95
- Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from the retrospective study of disease. *J. natl Cancer Inst.*, *22*, 719–748
- McGregor, D.H., Land, C.E., Choi, K., Tokuoka, S., Liu, P.I., Wakabayashi, T. & Beebe, G.W. (1977) Breast cancer incidence among atomic bomb survivors, Hiroshima and Nagasaki, 1950–69. *J. natl Cancer Inst.*, *59*, 799–811
- Miettinen, O.S. (1972) Components of the crude risk ratio. *Am. J. Epidemiol.*, *96*, 168–172
- Miettinen, O.S. (1974) Confounding and effect modification. *Am. J. Epidemiol.*, *100*, 350–353
- Miettinen, O.S. (1976) Stratification by a multivariate confounder score. *Am. J. Epidemiol.*, *104*, 609–620
- Newell, D.J. (1963) Misclassification in 2×2 tables. *Biometrics*, *19*, 187–188
- Peto, R. (1977) *Epidemiology, multi-stage models and short-term mutagenicity tests*. In: Hiatt, H.H., Watson, J.D. & Winston, J.A., eds, *Origins of Human Cancer*, Cold Spring Harbor, NY, Cold Spring Harbor Publications, pp. 1403–1428
- Pike, M.C., Anderson, J. & Day, N. (1979) Some insight into Miettinen's multivariate confounder score approach to case-control study analysis. *J. Epidemiol. Community Health*, *33*, 104–106
- Prentice, R. & Breslow, N.E. (1978) Retrospective studies and failure time models. *Biometrika*, *65*, 153–158

- Report of the Surgeon General (1964) *Smoking and Health*, Washington DC, US Government Printing Office (DHEW Publication No. 1103)
- Rogentine, G.N., Yankee, R.A., Gart, J.J., Nam, J. & Traconi, R.J. (1972) HLA antigens and disease: acute lymphocytic leukemia. *J. clin. Invest.*, 51, 2420–2428
- Rothman, K.J. (1974) Synergy and antagonism in cause-effect relationships. *Am. J. Epidemiol.*, 99, 385–388
- Rothman, K.J. (1976) The estimation of synergy or antagonism. *Am. J. Epidemiol.*, 103, 506–511
- Schlesselman, J.J. (1978) Assessing effects of confounding variables. *Am. J. Epidemiol.*, 108, 3–8
- Scott, R.C. (1978) The bias problem in sweep-and-smear analysis. *J. Am. stat. Assoc.*, 73, 714–718
- Seidmann, H., Lilis, R. & Selikoff, I.J. (1977) *Short-term asbestos exposure and delayed cancer risk*. In: Nieburgs, H.E., ed., *Prevention and Detection of Cancer*, Part 1: *Prevention*, Vol. I: *Etiology*, New York, Marcel Dekker, Inc., pp. 943–960
- Seigel, D.G. & Greenhouse, S.W. (1973) Validity in estimating relative risk in case-control studies. *J. chron. Dis.*, 26, 219–226
- Simon, D., Yen, S. & Cole, P. (1975) Coffee drinking and cancer of the lower urinary tract. *J. natl Cancer Inst.*, 54, 587–593
- Smith, P.G. & Doll, R. (1978) *Age- and time-dependent changes in the rates of radiation-induced cancers in patients with ankylosing spondylitis following a single course of x-ray treatment*. In: *Late Biological Effects of Ionizing Radiation*, Vol. I, Vienna, International Atomic Energy Agency, pp. 205–218
- Tulinius, H., Day, N.E., Johannesson, G., Bjarnason, O. & Gonzalez, M. (1978) Reproductive factors and risk for breast cancer in Iceland. *Int. J. Cancer*, 21, 724–730
- Whittemore, A.S. (1977a) *Epidemiologic implications of the multi-stage theory of carcinogenesis*. In: Whittemore, A.S., ed., *Environmental Health Quantitative Methods*, Philadelphia, SIAM, pp. 73–87
- Whittemore, A.S. (1977b) The age distribution of human cancers for carcinogenic exposures of varying intensity. *Am. J. Epidemiol.*, 106, 518–532
- Whittemore, A.S. (1978) The collapsibility of multi-dimensional contingency tables. *J. R. stat. Soc. B*, 40, 328–340
- Wynder, E.L. & Stellman, S.D. (1979) Impact of long-term filter cigarette usage on lung and larynx cancer risk: a case-control study. *J. natl Cancer Inst.*, 62, 471–479
- Wynder, E.L., Bross, I.J. & Feldman, R.M. (1957) A study of the etiological factors in cancer of the mouth. *Cancer*, 10, 1300–1323

LIST OF SYMBOLS – CHAPTER 3 (in order of appearance)

E	Exposure
C	Confounder
a	Number of exposed cases
b	Number of non-exposed cases
c	Number of exposed controls
d	Number of non-exposed controls

n_1	Number of cases
n_0	Number of controls
m_1	Number exposed
m_0	Number non-exposed
N	Total number
$\psi_p = ad/bc$	Pooled odds ratio relating E and disease
a_1	Number of exposed cases at 1st level of C
a_2	Number of exposed cases at 2nd level of C
(similarly for b, c, d, n_1 , n_0 , m_1 , m_0 , N)	
$\psi_1 = \frac{a_1 d_1}{b_1 c_1} =$	odds ratio relating E and disease at level 1, factor C
$\psi_2 = \frac{a_2 d_2}{b_2 c_2} =$	odds ratio relating E and disease at level 2, factor C
ψ	common value of ψ_1 and ψ_2 , assuming they are equal; within stratum odds ratio
ψ_c	odds ratio relating C and disease after stratification by E
p_1	proportion of exposed controls who are at level 1(+) of C
p_2	proportion of non-exposed controls who are at level 1(+) of C
$w = \psi_p/\psi$	confounding risk ratio
r_k	relative risk associated with level k of a confounder C
p_{1k}	proportion of exposed controls who are at level k of factor C
p_{2k}	proportion of non-exposed controls who are at level k of factor C
r_j^*	relative risk associated with j^{th} combination of levels of a confounder C
p_{1j}^*	proportion of exposed controls having the j^{th} combination of levels of factor C
p_{2j}^*	proportion of non-exposed controls having the j^{th} combination of levels of factor C
w^*	confounding risk ratio when K levels of factor C are grouped into J combinations (see equation 3.3)
M	ratio of controls to cases in a balanced design
Φ	probability that an exposed case or control is mistakenly classified as non-exposed
ϑ	probability that a non-exposed case or control is mistakenly classified as exposed
p_1	exposure probability for cases
p_0	exposure probability for controls