No cover image available

Critical Appraisal of Epidemiological Studies and Clinical Trials (3rd edn)

Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.001.0001

Published online: 01 September 2009 **Published in print:** 22 February 2007 **Online ISBN:** 9780191723865 **Print ISBN:** 9780198529552

CHAPTER

6 Confounding

J. Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.003.06 Pages 157–224

Published: February 2007

Abstract

Confounding is the most challenging issue in the interpretation of studies. This chapter is divided into three parts. The first part defines confounding and shows what effects it can produce. The second part deals with how confounding can be controlled. The third part considers some further applications of the logic of confounding. Self-test questions are provided at the end of the chapter.

Keywords: confounding, study interpretation, randomization, stratification, matching, multivariate methods, cohort studies, intervention studies

Subject: Public Health, Epidemiology

Collection: Oxford Scholarship Online

> *Thus it is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease; for the statistics shew that the classes which use these articles are bigger, healthier, and live longer than the class which never dreams of possessing such things.*
>
> —George Bernard Shaw: Preface to 'The Doctor's Dilemma'; 1906

Part 1. Confounding: definition and examples

Confounding is the most challenging issue in the interpretation of studies. This chapter is in three parts. In part 1 we will define confounding and show what effects it can produce. In part 2 we will deal with how confounding can be controlled. In part 3, we will consider some further applications of the logic of confounding.

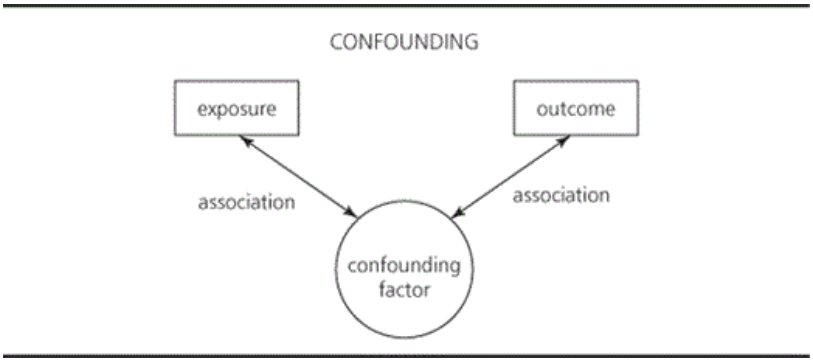
One of the central concepts of science in general, and biology in particular, is that of the tightly controlled experiment. In the classic laboratory experiment, the investigator uses laboratory rats that have been bred

under controlled conditions for many generations, are housed in identical physical environments, are fed and handled in the same way, and are then randomly allocated into the required groups. Observations are then made in a standardized manner, with the observer being 'blind' as to the allocation of the animal. The objective is to achieve a situation where the groups of animals differ in terms of only one factor, the exposure factor under consideration, and therefore there is no alternative but to assume that a difference in the measured outcome between the groups of animals is due either to that exposure factor or to chance. The randomization of animals from a common pool protects against there being other factors that differ between the groups, and the standardized and blind assessment procedure protects against bias in the observations of outcome.

In observational studies on humans such tight control is not possible, for scientific or more commonly for
p. 158 ethical or logistic reasons. Human subjects ↳ will differ from one another much more than the laboratory animals will, and we can only attempt to control a few aspects of their environment and activities. The challenge is to conduct studies of free-living human subjects that will still have a high degree of validity.

We have seen already that the results of a study, in terms of the differences between the groups being compared, may be due to any of four mechanisms: bias, confounding, chance, or causation.

Confounding is defined as (**Ex.** 6.1) *a distortion of an exposure–outcome association brought about by the association of another factor with both outcome and exposure*.

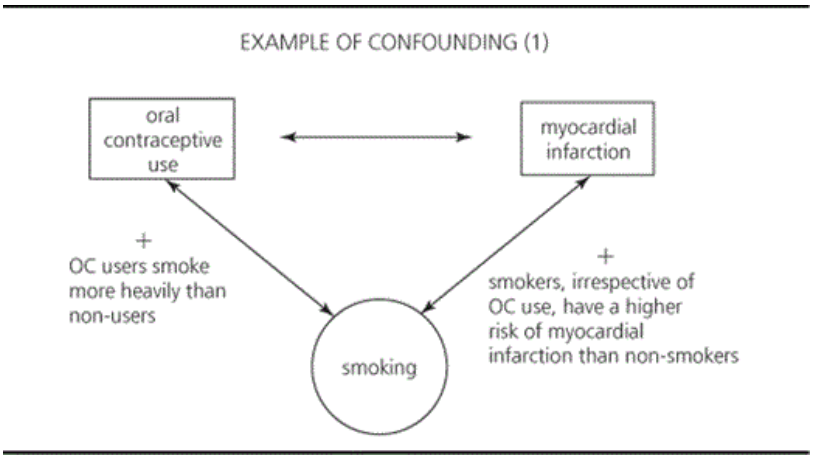


Ex. 6.1. Confounding, showing the two associations which are necessary for it to occur

To understand confounding, let us consider some situations intuitively. Suppose that we need to assess if there is a causal association between the use of oral contraceptives by women and the occurrence of myocardial infarction. There are two standard designs: a cohort study comparing oral contraceptive users with non-users, and a case–control study comparing women with myocardial infarction with an unaffected comparison group.

The issue we are to deal with now is: can an observed association between oral contraceptive use and myocardial infarction be influenced by differences between the two groups of women in terms of other factors? Consider first the issue of smoking. There is ample evidence that people who smoke have an increased risk of myocardial infarction. There is also evidence, in some communities at least, that women who use oral contraceptives smoke more than women who do not. Now consider the effect of these two associations on the results of these studies. Consider the situation where the null hypothesis is in fact the truth, i.e. there is no
p. 159 causal association between oral contraceptive use ↳ and myocardial infarction. In the cohort study, because the oral contraceptive users smoke more than the non-users, they will have a higher risk, shown by a higher incidence rate, of myocardial infarction. In the case–control study, because smoking is a causal factor for myocardial infarction, the prevalence of smoking will be greater in the myocardial infarction patients than in the comparison patients; and, because smoking is associated with oral contraceptive use, oral contraceptive use will also be more common in the myocardial infarction patients.

Thus both studies will give a result suggesting a positive relationship between oral contraceptive use and myocardial infarction, even if there is no true causal relationship (**Ex.** 6.2). If the true situation is that oral contraceptive use increases the risk of myocardial infarction, this confounding effect will mean that the true effect is overestimated. If the true situation is that oral contraceptive use decreases the risk of myocardial infarction, this confounding effect will mean that the true protective effect will be underestimated or not detected.

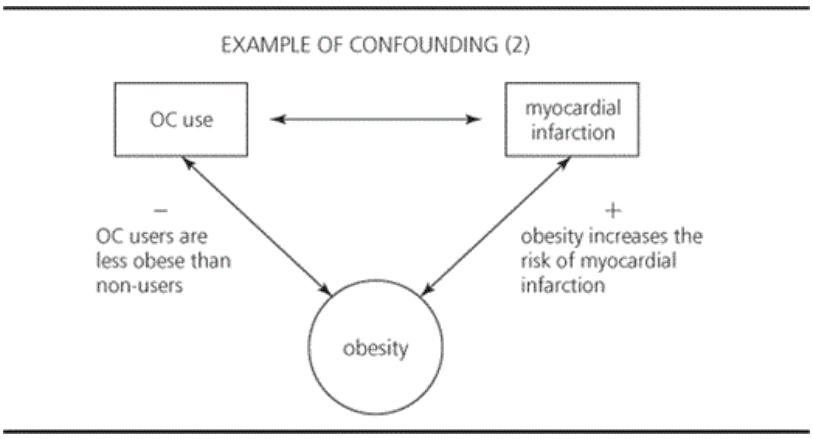


Ex. 6.2. An example of confounding: any comparison of the risks of myocardial infarction in oral contraceptive (OC) users and in non-users (cohort design), and any comparison of the prevalence of past OC use in myocardial infarction patients and in unaffected subjects (case–control design), will be influenced by the associations of both OC use and myocardial infarction with smoking, which is a confounding factor. The measured association between OC use and myocardial infarction will overestimate the true association

In this situation smoking is a confounding factor. Confounding is produced by the two simultaneous and
p. 160 independent properties: smoking is associated ↳ with the outcome in this study, and independently smoking is associated with the exposure.

The definition of confounding involves a definition of the study hypothesis, because it may be that smoking and oral contraceptives are both causal factors for myocardial infarction. Therefore if we were studying the relationship between smoking and myocardial infarction in women, we should have to consider oral contraceptive use as a potential confounding factor.

The effects of a confounding factor can be in either direction. In the situation given, the exposed group (oral contraceptive users) has a *higher* prevalence of smoking, and smoking is associated with an *increase* in risk of myocardial infarction. The net result of this confounding will be to give an apparent *excess* risk of myocardial infarction in the oral contraceptive users. In another situation, consider the relationship of oral contraceptive use to myocardial infarction in women and the effect of obesity as a confounding factor (**Ex.** 6.3). Suppose that oral contraceptive users are *less* obese than non-users, but that obesity gives an *increased* risk of myocardial infarction. In this situation the 'exposed' group of oral contraceptive users will be less obese than the group of non-users, and because of this their risk of myocardial infarction will be *reduced*. If the null hypothesis of no
p. 161 association between oral contraception and myocardial infarction is true, the study will give ↳ a spurious indication of a protective effect. If there is a real increase in the risk of myocardial infarction in oral contraceptive users, the study will underestimate this risk, and may not show it at all, if it fails to take into account the counteracting effect of the difference in obesity between contraceptive users and non-users.



Ex. 6.3. Negative confounding: if obesity increases the risk of myocardial infarction, and oral contraceptive (OC) users are less obese than non-users, the measured association between OC use and myocardial infarction will underestimate the true association

How do we know if the two associations critical to confounding exist? The issue is not whether the associations exist in general, for example whether it is true in general that oral contraceptive users are less obese than non-users. That would be a difficult claim to substantiate, as the relationship is likely to vary between women of different ages, in different countries, and so on. That is not important. The crucial issue is: does the association exist within the study population, within the data used in the analysis? Thus, if within the data set given by a particular study it is true that oral contraceptive users are less obese than non-users, and that obese subjects have a higher incidence of myocardial infarction, then obesity will be a confounding factor in the relationship between oral contraceptive use and myocardial infarction. The only other proviso is that the obesity–oral contraceptive and obesity–infarction associations must apply independently from the oral contraceptive–infarction relationship, i.e. obesity must be associated with oral contraceptive use even in women without infarction, and obesity must be related to infarction even in women who do not use oral contraceptives. This may seem difficult logic, but should be clearer after some examples are presented.

Therefore a factor is a confounding factor only when the *two* associations exist—when *the factor is associated with both the exposure and the outcome* under assessment. Oral contraceptive users and non-users may differ in terms of many other factors; for example, they may differ in their exposure to hair dyes, with oral contraceptive users more frequently using hair dyes. Do we have to consider hair dye use as a confounder in assessing the study? The answer is, only if hair dye use is itself related to myocardial infarction. If we have asked questions about hair dye use in our study, we can assess this by looking at the women who are not exposed to oral contraceptives, and within that subpopulation see if there is an association between hair dye use and myocardial infarction. If there is not, there is no need to consider hair dye use as a confounder. Similarly, there may be factors related to the outcome but not to the exposure. For example, the risk of myocardial infarction is increased in women who have certain types of familial hypercholesterolaemia. This will be a confounding factor only if in the study in question the prevalence of hypercholesterolaemia differs between oral contraceptive users and non-users.

p. 162

Confounding in cohort and intervention studies

Because an understanding of confounding is so important, we shall look at a number of simple examples, using both hypothetical and real data. Consider the hypothetical example in **Ex.** 6.4 which shows the results of a cohort study in which subjects with low exercise levels are compared with subjects with high exercise levels; the outcome under investigation is the incidence of myocardial infarction in a given follow-up period.

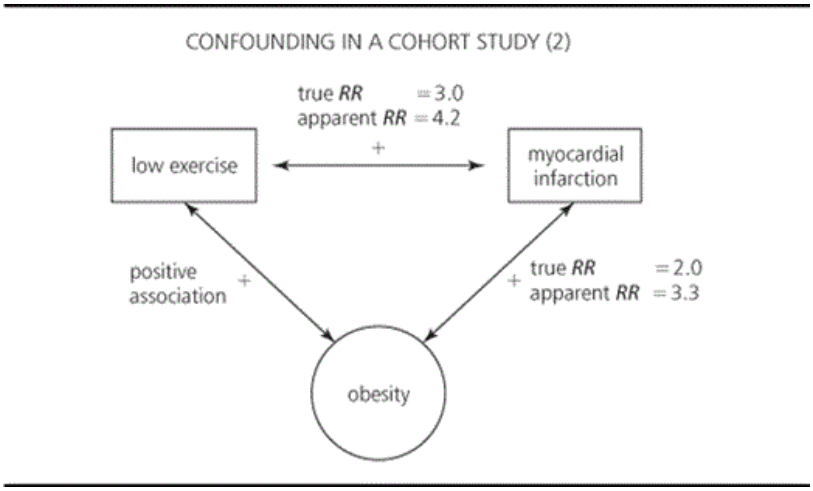
CONFOUNDING IN A COHORT STUDY (1)

	Myocardial infarctions	Person-years	Incidence/1000
Table A: all subjects (n = 8000 person-years)			
Low exercise	105	4000	26.25
High exercise	25	4000	6.25
Relative risk = 26.25/6.25 = 4.2			
Subtable B_1: obese subjects (n = 4000)			
Low exercise	90	3000	30.0
High exercise	10	1000	10.0
Relative risk = 3.0			
Subtable B_2: non-obese subjects (n = 4000)			
Low exercise	15	1000	15.0
High exercise	15	3000	5.0
Relative risk = 3.0			

Ex. 6.4. Confounding: a cohort study assessing the association between the incidence of myocardial infarction and exercise, where obesity is a confounding factor (hypothetical data)

Table A shows the simplest form of the results, showing a strong association with low exercise people having a relative risk of 4.2 compared with high exercise people. However, let us assume that the subjects vary in obesity, and that obesity and exercise are related; obesity is less common in the high exercise subjects than in the low exercise subjects. Subtables B_1 and B_2 show the results in an identical format to the first table, separately for obese subjects and non-obese subjects. Subtable B_1 shows that the relative risk for low compared with high exercise in subjects who are obese is 3.0. Subtable B_2 shows that the relative risk for low compared with high exercise in subjects who are not obese is also 3.0. Therefore, irrespective of obesity, the best estimate of the effect of exercise is clearly 3.0. Why then did we get the result of 4.2 in the first table for all subjects? The
p. 163 reason ↳ is that obesity is a confounding factor. Obesity is itself a risk factor for myocardial infarction, and this can be seen by comparing the risks for obese subjects for a given level of exercise (Subtable B_1) with those for non-obese subjects with the same level of exercise (Subtable B_2). For low exercise subjects, the incidence rates per 1000 are 30.0 in obese and 15.0 in non-obese subjects; for high exercise subjects, the corresponding rates are 10.0 and 5.0. Moreover, obesity is related to exercise level, as comparison of the subtables shows that low exercise subjects are much more obese in terms of their distribution by person-years of experience; for the low exercise group, 75 per cent of the person-years apply to obese subjects, while for the high exercise group the proportion is 25 per cent.

The relationships between these factors, low exercise, obesity, and the outcome of myocardial infarction, are shown diagrammatically in **Ex.** 6.5. Low exercise is a risk factor for myocardial infarction, with a true relative risk of 3.0. Obesity is also a risk factor for myocardial infarction, with a relative risk of 2.0; this result is derived from a comparison of Subtables B_1 and B_2. However, because low exercise and obesity are positively related to each other, the apparent relative risk of low exercise, based on simply comparing all low exercise subjects with all high exercise subjects, is 4.2. Further, we can add the data in Subtables B_1 and B_2 to compare all obese subjects with all non-obese subjects; we obtain the apparent risk ratio of 3.3 for the crude relationship between obesity and myocardial infarction. Thus in this situation there are two independent risk factors for myocardial
p. 164 infarction, low exercise and obesity, which are positively correlated with each other; therefore each acts ↳ as a confounding factor when the relationship of the other to myocardial infarction is assessed.



Ex. 6.5. Confounding: the associations which exist in Ex. 6.4. *RR* = relative risk

Consider now a real example of rather simple confounding, shown in **Ex.** 6.6. This is derived from a 1986 paper in the *British Medical Journal*, which amongst other comparisons (a third method, lithotripsy, was also assessed) compared the success rate for two different surgical procedures in the treatment of renal calculi [1]. The upper table shows the results as they were described in the summary of the paper. For each surgical technique, open surgery and percutaneous nephrolithotomy, 350 patients were assessed, and the success rates were 78 per cent with open surgery and 83 per cent with percutaneous nephrolithotomy. In this study patients were categorized into those who had stones of less than 2 cm in diameter, and those with larger stones. For patients with small stones, the success rate of open surgery was better than that of the other technique: 93 per cent compared to 87 per cent. For patients with larger stones, open surgery also had a higher success rate: 73 per cent compared to 69 per cent. Thus, for either of the two groups of patients, open surgery gave better success rates. An erroneous
p. 165 impression of a lower success rate is created from the pooled data, because open surgery was used much ↳ more often on patients with large stones, and those patients had a lower success rate irrespective of the technique used. (As an aside, when this error was pointed out and a more sophisticated analysis suggested, the authors rejected this as '... it would only confound the clinicians' [2]. In fact, no complex analysis is needed, just a cross tabulation as in Ex. 6.6.)

CONFOUNDING: TREATMENT OF RENAL CALCULI

	Successes	Failures	Total patients	Successes (%)
All stones (n = 700)				
open surgery	273	77	350	78
percutaneous nephrolithotomy	289	61	350	83
Stones < 2 cm (n = 357)				
open surgery	81	6	87	93
percutaneous nephrolithotomy	234	36	270	87
Stones ≥ 2 cm (n = 343)				
open surgery	192	71	263	73
percutaneous nephrolithotomy	55	25	80	69

Ex. 6.6. Confounding: a comparison of two surgical methods of treating renal calculi, showing success rates (percentage of patients with no stones at 3 months after treatment). The summary of this paper states 'success was achieved in 273 (78 per cent) of patients after open surgery, 289 (83 per cent) after percutaneous nephrolithotomy, …'. However, in fact the success rates for open surgery are higher, not lower, than those for the percutaneous technique. The main result of this paper concerns a third method, extracorporeal shock-wave lithotripsy, which was followed by higher success rates than those shown. From Charig *et al*. [1]

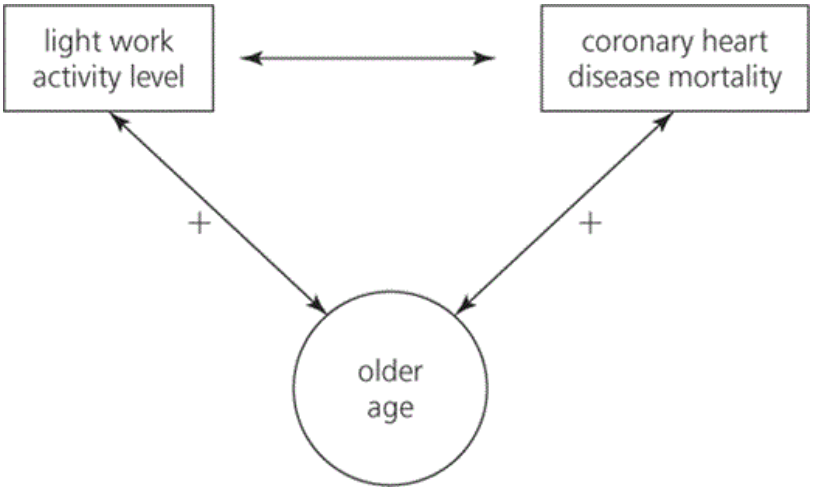
In this situation, the confounding factor has produced a reversal in the direction of a relationship: while percutanoeus nephrolithotomy in fact had the higher success rate, the uncontrolled confounding by size of stone gave results showing a higher success rate with open surgery. While there is nothing different in this than in other confounding effects, such a reversal of effect is sometimes called Simpson's paradox. Simpson was a statistician who discussed confounding with a hypothetical example in a 1951 paper [3]. He showed how confounding would occur unless the factor involved was independent of either the exposure or the outcome, but he did not show or emphasize a reversal of effect.

Often the confounding factor will have more than two categories. A further example of real data (**Ex.** 6.7) shows the relationship between physical activity and mortality from coronary heart disease in the prospective study of longshoremen (dockworkers) in California, noted in Chapter 3 [4]. Table A in **Ex.** 6.7 shows the total data comparing light or moderate exercise level workers with heavy exercise level workers, and shows a relative risk of 3.4 in the light exercise group. However, as one might predict, there was considerable confounding by age in this study. Workers doing the lighter physical work tended to be older than those doing the heavier work. Therefore their high relative risk could have been because they were older, rather than a direct effect of their lower exercise levels. Thus, in Table B the results are subdivided into four age groups. The relative risks for each age group range from 1.1 to 2.0. We can see that the true effect of exercise averaged over all workers must be some figure between these numbers, and cannot be as high as the observed crude relative risk of 3.4, which is produced partly by the difference in age distribution. It is not intuitively obvious what the best single estimate of the effect of exercise would be; that will be discussed later.

CONFOUNDING IN A COHORT STUDY

Activity level	Deaths	Man-years	Rate/10 000	Relative risk
Table A. All ages				
Light or moderate	532	65 000	81.8	3.4
Heavy	66	27 700	23.8	1.0 (referent)
Table B. Age 35–44				
Light or moderate	3	5900	5.1	1.1
Heavy	4	8300	4.8	
Age 45–54				
Light or moderate	62	17 600	35.2	1.9
Heavy	20	11 000	18.2	
Age 55–64				
Light or moderate	183	23 700	77.2	1.7
Heavy	34	7400	45.9	
Age 65–74				
Light or moderate	284	17 800	159.6	2.0
Heavy	8	1000	80.0	

Ex. 6.7. Confounding: data from a cohort study of mortality from coronary heart disease and exercise. Confounding by age distorts the association between physical activity and mortality from heart disease. In Table A, a comparison of men with light or moderate physical activity with those with heavy activity gives a relative risk of 3.4. Table B shows that (1) mortality rises with age and (2) the proportion of men doing light or moderate work rises with age. Age is a confounding factor, and its effect gives an increase in the observed relative risk between light activity and CHD mortality. This excess is shown as the relative risks within each of four age bands are all much lower than the crude estimate of 3.4. From Paffenbarger and Hale [4]. See also Chapter 3, p. 60. The situation can be represented as:



Confounding in case–control studies

The logic of confounding in case–control studies is identical to that in cohort studies, but the arithmetic is slightly different. **Ex.** 6.8 shows a simple hypothetical example of a case–control study comparing patients with skin cancer with controls, the exposure of interest being eye colour in two categories, blue and brown.

CONFOUNDING IN A CASE–CONTROL STUDY (1)

	Cases	Controls
Table A. All subjects (n = 1000)		
Blue eyes	257	200
Brown eyes	243	300
Total	500	500
Odds ratio = (257 × 300)/(243 × 200) = 1.59		
Subtable B_1 Dark hair colour (n = 550)		
Blue eyes	57	100
Brown eyes	143	250
Total	200	350
Odds ratio = (57 × 250)/(143 × 100) = 1.00		
Subtable B_2 Light hair colour (n = 450)		
Blue eyes	200	100
Brown eyes	100	50
Total	300	150
Odds ratio = (200 × 50)/(100 × 100) = 1.00		

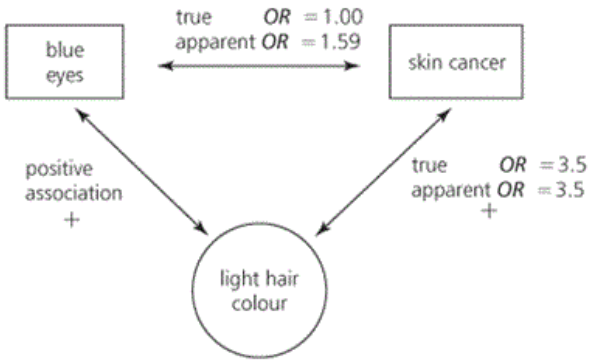
Ex. 6.8. Confounding in a case–control study: a hypothetical example examining the relationship of eye colour to skin cancer

p. 167 In Table A, the results for 500 cases and 500 controls are shown, giving an odds ratio of 1.59 for blue compared with brown eyes. However, hair colour has been shown to be a risk factor for skin cancer, so Subtables B_1 and B_2 show the association between disease and eye colour for subjects with dark or light hair colour separately. For those with dark hair the odds ratio is 1.0, showing no association, and for those with light hair it is also 1.0. Clearly the best estimate of the association of eye colour with skin cancer is an odds ratio of 1.0, showing no association. We have an apparent association between blue eyes and skin cancer, with an odds ratio of 1.59, although there is no true association. The apparent excess risk in individuals with blue eyes is because of a positive association between blue eyes and light skin colour. This association can be seen by comparing Subtables B_1 and B_2 in terms of the control subjects: the proportion of control subjects with light hair colour who have blue eyes is 67 per cent (100/150), while the proportion of those with dark hair who have blue eyes is only 29 per cent (100/350)

Subtables B_1 and B_2 can also be used to look at the relationship between hair colour and skin cancer. The

p. 168 apparent odds ratio, comparing all light-haired ↳ with all dark-haired subjects, is (300 × 350)/(200 × 150) = 3.5. The true odds ratio is obtained by examining hair colour within the categories of eye colour. For subjects with blue eyes, the odds ratio for the relationship between light hair colour and skin cancer is (200 × 100)/(100 × 57) = 3.5; for subjects with brown eyes it is (100 × 250)/(50 × 143) = 3.5. Therefore adjusting for eye colour makes no difference to the association between hair colour and skin cancer, showing that eye colour is not a confounder in the relationship between hair colour and skin cancer. The reason eye colour is not a confounder is that although it is associated with hair colour, it is not itself a risk factor for skin cancer, as shown by its true odds ratio with skin cancer being 1.0. The associations are shown diagrammatically in **Ex.** 6.9. In fact, the calculation of odds ratio for hair colour within the categories of eye colour was unnecessary. Given that there is no association between eye colour and skin cancer, once hair colour is controlled, we can deduce that eye colour cannot be a confounder because it does not fit the definition of a confounder; it is not associated with the outcome under study. Therefore the crude odds ratio relating hair colour to skin cancer will not be affected by controlling for eye colour. A further example of confounding in a case–control study, based on real data, will be shown in Ex. 6.23.

CONFOUNDING IN A CASE CONTROL STUDY (2)



Ex. 6.9. Confounding: a diagram showing the relationships present in the data given in Ex. 6.8

STRATIFICATION IN A CASE–CONTROL STUDY

	Cases	Controls
Table A: all subjects (n = 924)		
Exposed: severe sunburn	136	98
Unexposed: mild sunburn	343	347
Total	479	445
Odds ratio = (136 × 347)/(98 × 343) = 1.40		
Subtable B_1: subjects who sunburn easily (n = 595)		
Severe sunburn	119	76
Mild sunburn	227	173
Total	346	249
Odds ratio = (119 × 173)/(76 × 227) = 1.19		
Subtable B_2: subjects who do not sunburn easily (n = 329)		
Severe sunburn	17	22
Mild sunburn	116	174
Total	133	196
Odds ratio = (17 × 174)/(22 × 116) = 1.16		

Mantel-Haenszel estimate of odds ratio

$$= \frac{119 \times 173/595 + 17 \times 174/329}{76 \times 227/595 + 22 \times 116/329}$$
$$= 1.19$$

Ex. 6.23. Stratified analysis of a case–control study comparing patients with malignant melanoma with community controls with regard to history of sunburn, and adjusting for tendency to sunburn. Simplified from Elwood *et al*. [22]

Part 2. Methods for the control of confounding

Having understood what confounding is, we can consider the methods available to deal with it. There are only
p. 169 five methods available, as shown ↳ in **Ex.** 6.10. In the *design* of a study we can *restrict* the participation in the study to certain individuals; we can decide to *match* individuals in the comparison group to individuals in the group of interest; and we may have the option of doing a *randomized* intervention study. Irrespective of what has been done at the design stage, when we come to the *analysis* we may again use *restriction* to certain individuals in the data set; we may divide the data into subgroups by categories of the confounding factor, which is the process known as *stratification*; or we may use *multivariate* mathematical methods to take into account the effect of more than one confounding factor simultaneously. In practice a combination of these methods is usually used. We will discuss each of these five methods.

METHODS OF CONTROLLING CONFOUNDING	
In the design of the study	restriction matching randomization
In the analysis of the study	restriction stratification multivariate methods

Ex. 6.10. The methods of controlling confounding: one or more may be used

Control of confounding: restriction

Let us go back to our example of a study of oral contraception and myocardial infarction, and consider how we could avoid being misled by the effect of smoking. One way would be to include only women who had never smoked in the study population. We could do this with either a cohort or a case–control design, and we refer to this method as 'restriction'. It is clearly an effective method, as it leaves no possibility of confounding, but obviously the disadvantage is that the study then becomes specific to the relationship between oral contraceptive use and myocardial infarction in non-smokers, and we cannot generalize the study beyond that target population. (Of course, this restriction only controls for active smoking by the woman herself, not for exposure to smoking by other people. Indeed, a study restricted to never-smokers would be a good design to assess the effects of passive smoking, as the major confounding factor of active smoking is removed).

Suppose instead we do the entire study on women who smoke or have smoked. Can there still be confounding
p. 170 by smoking in comparing, for example, ↳ a cohort of oral contraceptive users who smoke with non-oral contraceptive users who smoke? The answer is that there is still potential for confounding. The fact that all the women in the study smoke does not mean that they all smoke the same amount, and if smoking has a dose–response relationship with myocardial infarction, and if the amount or duration of smoking is different in users and non-users, there is still potential for confounding. However, if we restrict the entire study to women who had all smoked, for example, 10–20 cigarettes per day for 5–10 years, the extra specification in the design will reduce the extent of any confounding. However, such a precise restriction would be cumbersome to apply in a practical study.

All studies involve some restriction, if only for practical reasons. The source and eligible populations will be restricted in terms of calendar time, geographical location, and frequently other factors such as age. Restriction should be considered if it is clear that within the study there may be relatively small groups of individuals whose results may be appreciably different from those of the main study population, and for whom the study is unlikely to provide much useful information because of the small numbers. A frequent situation is that of a

racial or ethnic group which contributes only a small proportion of subjects. It may have substantially different outcome rates and exposure histories, but there is little value in including such subjects if their total number is likely to be too small for independent consideration.

Randomization

We have dealt with restriction first because some degree of restriction applies to all studies. We deal with randomization next because it has many advantages over the other methods, and therefore there is a major distinction between randomized and non-randomized studies, both in design and evaluation. If a randomized intervention study is feasible, it has many advantages and should be considered first. However, it is only relevant for certain situations: prospective intervention studies assessing the effects of an ethical, practical, and acceptable intervention which is potentially beneficial and not likely to be harmful. Thus while of prime importance in assessing the effectiveness of clinical or public health interventions, randomization is not relevant to most issues regarding the causes of disease and cannot be applied to retrospective studies.

The principle of randomization is that from a pool of study entrants, subjects are randomly assigned to each of the intervention and non-intervention groups. The definition of *random* is that each subject in the study has
p. 171 the same chance of being allocated to any particular group, and that the chance ↳ of a particular individual being allocated to one group is not influenced by the allocation of any other individual. It is not simply an unsystematic or haphazard assignment. Methods based on an apparently random process, like tossing a coin, will work in principle, but in practice are too open to the possibility of conscious or unconscious manipulation. Randomization is normally done by reference to numbers generated to be completely random obtained either from a computer program or from a table of such numbers available in a standard statistical text.

The allocation sequence must be random, and the integrity of the process must be protected, and seen to be protected. Each study participant, and the staff working with them, should complete all the study entry requirements before their assignment to an intervention group is made. Subjects should be invited to participate in the study, be confirmed as eligible according to the study protocol, and should give informed consent; and their participation should be recorded. Only then should their allocation to a specific intervention group be given. The assignment and the decision about eligibility should not be changed after the assignment has been revealed. To achieve this, the randomization process is best administered independently from the recruitment of the subjects. The randomization should be done by someone who is not involved in subject recruitment, such as a statistician or pharmacist; in large trials it is often done by staff in an independent office. A record should be kept of all subjects assessed and considered eligible for trial entry, showing whether they gave consent to the trial, and then whether they were randomized. All subjects randomized should be considered as trial participants from the time of randomization.

The essential logic is that random allocation makes it *likely* that the two groups created will be similar with respect to any particular variable. The essential limitation of randomization is that it is a method based on probability. Therefore its chances of success will be great only if substantial numbers of subjects are used, and we can never be *certain* that randomization will provide equivalent groups. In the extreme, randomization in a study of only two individuals may protect against investigator bias in the assignment, but does not reduce the differences between the two subjects. In a large randomized study, it is highly likely that the groups created by randomization will be comparable with respect to specific factors. However, if the numbers in each group are relatively small (a reasonable guideline might be less than 100), then it is quite likely that purely by chance the groups will still vary. If they vary in terms of a factor which is strongly related to the outcome in the study, that factor will be a confounding factor, and we must deal with that in the analysis. The practical message is that
p. 172 randomization is a valuable technique, which with reasonable ↳ numbers of subjects should work in most situations; but we should not assume that simply because randomization has been used, the groups being

compared cannot differ in terms of any confounding factor. Just as in any other study, data on the factors likely to be the main confounders should be used to compare the groups to make sure they are similar. If they are not, even in a randomized study, other methods of analysis such as stratification or multivariate methods may also be used to take account of any differences of confounding factors.

The value of randomization

Exhibit 6.11 shows the value of randomization. In a study previously mentioned in Chapter 2, the provision of iron rather than aluminum cooking pots to families was assessed as a method of reducing anaemia and improving growth in children in Ethiopia [5]. A randomized controlled trial was used, the randomization being by household, with one child per household participating. It might be expected that any comparison other than by randomization would result in the families using iron pots and those using aluminium pots differing in terms of a wide range of characteristics, which could be themselves related to child development. Exhibit 6.11
p. 173 shows that the groups created by the ↳ randomization process are very similar in terms of the age, sex, height, and weight of the children, their recent medical history, characteristics of the mother, and household access to clean water and sanitation.

RANDOMIZATION TO PRODUCE EQUIVALENT GROUPS

	Randomized to iron pots (n = 195)	Randomized to aluminium pots (n = 212)
Age (months, mean and std deviation)	31.3 (14.6)	30.5 (15.7)
Male/female (numbers)	99/96	106/106
Weight (kg, mean and SD)	11.6 (2.3)	11.9 (2.3)
Length (cm, mean and SD)	87.2 (8.5)	88.0 (8.7)
Ill in week preceding study	57 (29%)	70 (33%)
Diarrhoea in week preceding study	35 (18%)	40 (19%)
Mother literate	72 (37%)	68 (32%)
Mother ill in last 7 days	50 (26%)	49 (23%)
Access to clean water	148 (76%)	157 (74%)
Adequate sanitation	57 (29%)	72 (34%)

Ex. 6.11. The benefits of randomization: in a trial in Ethiopia assessing whether the use of iron pots or aluminum pots affected anaemia and weight gain in children, households were randomly allocated to receive either iron or aluminum pots, and one child per household participated in the study. These data show the similarity of the two groups in regard to characteristics at the time of randomization. SD = standard deviation. From Adish *et al*. [5]

We can look at these two groups of subjects and consider whether, if they were subsequently treated in an identical manner, we should expect their outcomes to be the same; there is little in the table that would suggest otherwise. Randomization is the simplest way to achieve such equivalent groups. In principle, we could have a design where the first family seen was allocated an iron pot, and was matched to another family on the features shown in Ex. 6.11 which would be given an aluminium pot, but such a study would be difficult or impossible. The results of this study were that children in households using iron pots had a greater rise in haemoglobin concentration, and gained more in both weight and height over a 12-month period, and it was concluded that the provision of iron cooking pots may be a useful way to prevent iron deficiency anaemia in similar less developed countries [5].

There is one advantage of randomization that is shared by no other technique. This is that randomization, given reasonably large numbers of subjects, is likely to produce groups that are similar even with respect to variables that we have not anticipated, defined, or measured. Suppose that after the study just described is completed, evidence appears that an infection common in this community is a strong predictor of childhood anaemia. The study would have been better if that infection had been assessed, and if it were shown that the

groups were similar in terms of it. However, even without those data, as the original study was randomized and had adequate numbers, we can be reasonably sure that the distribution of the two groups in terms of this unmeasured factor would have been similar.

The limits of randomization: pre-stratification

The amount of confounding produced by a factor depends on the strength of its association with the outcome, and the strength of its association with the exposure. Consider a comparison of two treatments for lung cancer. The outcome, mortality, will vary greatly with the extent (stage) of the disease; because this association (stage-outcome) is very strong, stage may have a major confounding effect even if the difference in stage distribution between the treatment groups is small. It is not appropriate to assess confounding by applying a statistical test to compare the stage distribution in the two treatment groups. Stage could be an important confounder even if the difference in stage between the two groups is not statistically significant. Papers describing randomized trials should have a table such as Ex. 6.11 showing the distribution of relevant factors in the randomized groups,
p. 174 but statistical tests should ↳ not be used. This is recommended in the CONSORT statement [6], which was discussed in Chapter 4 (Ex. 4.8, p. 96).

It follows from this that where some major confounding factors can be predicted in advance, it may be better not to rely only on the randomization procedure to produce similar groups. A more reliable procedure is to group the eligible subjects within categories of the strong confounder, and randomize within these categories. Thus in the above example we could classify all study entrants by stage of disease, and randomize within each stage, thus ensuring that the stage distribution of the treatment groups will be virtually identical. This is a combination of randomization and stratification, and is sometimes referred to as *randomization within blocks*, or *pre-stratification*. For example, in a trial of perineal pain relief after childbirth, women were randomized within four strata, determined by parity (first birth and later births) and mode of birth (spontaneous and instrumental) as these factors were important with regard to perineal trauma and pain [7]. However, randomized trials are often difficult and time consuming in practice; simple designs have great advantages in clinical studies, and pre-stratification should be used cautiously. It is often used to randomize within centres in multicentre studies, ensuring that each centre treats similar numbers of subjects on each of the alternative therapies.

Difficulties with randomized studies

A randomized trial makes heavy demands on participants, health care professionals, and those involved indirectly such as service managers and support staff. Randomization is often difficult to use for reasons of logistics or informed consent. For example, Cook *et al*. [8] describe a randomized trial comparing two drugs for the management of heroin withdrawal. The trial required support from a charitable foundation and from the manufacturers of the drugs, logistical support in hospital bed capacity and laboratory services, and the support of hospital staff and the patients themselves. All these presented difficulties so that the trial could not be completed.

The design of trials has to balance the ideal scientific design with practical considerations. For example, to evaluate a new method of encouraging smoking cessation in pregnant women, it may be administratively much easier to offer the new programme to all women in a particular clinic, and compare them with women in a different clinic, than to allocate women randomly in each clinic. Such a systematic allocation method is weaker than randomization, as there is a greater chance of the groups chosen differing in terms of relevant factors, and analytical methods need to take account of the group allocation (as will be discussed later).

p. 175 Questions of informed consent may be a critical influence on study design. In a randomized trial in Australian general practice assessing a new method of assessing skin lesions, the choice was between a study design in

which individual patients were randomized, and one where general practices were randomized. If patients within each practice were randomized, the general practitioner would need to discuss both the new and the conventional assessment systems and the randomized design with each potential participant, and obtain their written consent. If practices were randomized, the general practitioners would need to give their own consent to being involved in the study, but in their opinion, and that of the ethics committees, formal consent from individual patients would not be required as both management options could be used in ordinary practice. While this ethical distinction is debatable, the result was that a study involving individual randomization was impractical because of the time it would require busy general practitioners to discuss the randomization process with each patient. The trial based on randomization of practices was carried out, with enthusiastic support from practitioners and patients [9].

Randomization achieves its objectives by a random process. The principle is that it is *likely* that the groups produced by randomization will be equivalent. However, some differences between the groups will remain, and
p. 176 on some occasions these differences may be substantial and important. **Exhibit** 6.12 shows a ↳ table of baseline characteristics for subjects in an important randomized study of the treatment of diabetes, comparing those randomized to receive diet and tolbutamide (an oral glucose-lowering agent) with those randomized to receive diet plus a placebo; the outcome of interest was subsequent deaths, of which many were from cardiovascular disease [10]. Comparing these two groups shows that the patients randomized to receive tolbutamide were older and more frequently had a history of digitalis use or angina, and higher proportions had an electrocardiograph abnormality, high cholesterol levels, high glucose levels, increased relative body weight, and arterial calcification assessed by a radiograph of the lower limb. On the other hand, there was a lower proportion with a history of hypertension. With these data, we cannot be confident that if the two treatments used had identical effects, the two groups would show the same results in terms of subsequent mortality. Several of these factors could have considerable effects on subsequent mortality, and the differences between the groups appear substantial.

RANDOMIZED GROUPS MAY DIFFER

	Randomized to diet + tolbutamide $n = 204$ (% of subjects)	Randomized to diet + placebo $n = 205$ (% of subjects)
Age > 55	48.0	41.5
Digitalis use	7.6	4.5
Angina	7.0	5.0
ECG abnormality	4.0	3.0
Cholesterol > 300 mg/100 ml	15.1	8.6
Fasting glucose > 110 mg/100 ml	72.1	63.5
Relative body weight > 1.25	58.8	52.7
Arterial calcification	19.7	14.3
Hypertension	30.2	36.8

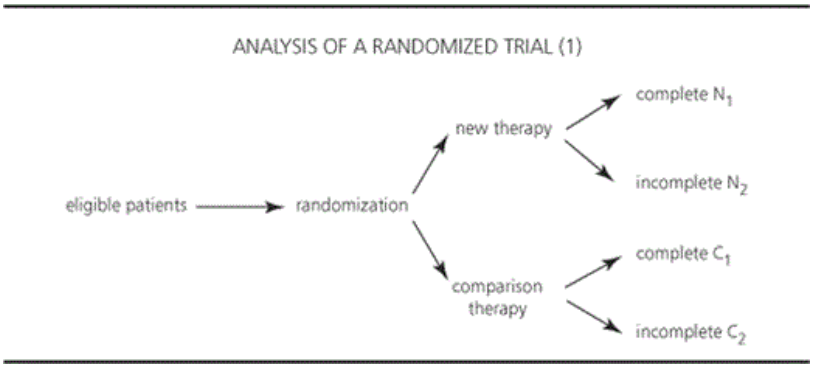
Ex. 6.12. The limits of randomization: this study compared four regimes for the management of diabetes; here the subjects randomized to diet + tolbutamide (an oral glucose-lowering agent) are compared with those randomized to diet + placebo. From University Group Diabetes Program [10]

Sometimes this is loosely referred to as a 'failure of randomization', but that is an inappropriate term. The randomization process has been carried out correctly, but being a probabilistic technique it does not guarantee that the groups will be similar in terms of all factors. Statistical tests show that all but one (cholesterol levels) of these differences are not statistically significant at the conventional 5 per cent level, i.e. differences of this or greater magnitude would be expected to occur on more than 5 per cent of occasions. However, this is not the relevant issue. The relevant issue is whether the differences are sufficiently large to influence the subsequent outcome rates in the two groups; a small difference in a factor which is strongly related to outcome will be

important. Thus in Ex. 6.12 we have a randomized study in which there are important differences between the groups being compared. The results of this study showed that total deaths, and particularly deaths from cardiovascular disease, were substantially higher in the group treated with tolbutamide than in the placebo group. The crucial question is whether this difference in mortality can be attributed to the tolbutamide, or whether it is due to the other factors that differ between the two groups. As will be shown later in this chapter, other analytical techniques can be used to address this question.

Analysis of randomized trials

In most randomized trials not all individuals complete the treatment to which they have been randomized. Some patients may start the treatment but not complete it; they may decide to discontinue, either for reasons
p. 177 related to the treatment (e.g. side effects) or for other reasons (e.g. change of residence), ↳ or their clinical situation may change so that a change in treatment is indicated. Some patients may be randomized but not even commence treatment. **Exhibit** 6.13 represents a clinical trial in which patients are randomized into two treatment groups, but only some of the patients complete the course of treatment offered. The question is: to assess the effect of the new therapy, which groups of patients should be compared?

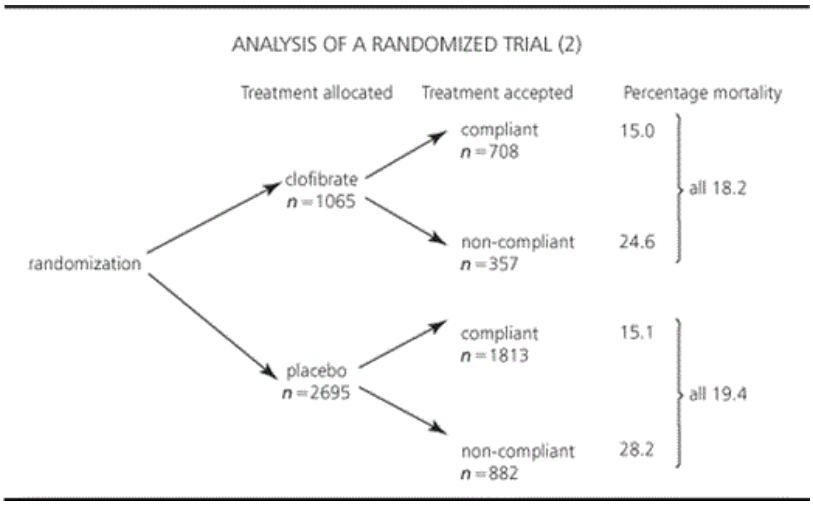


Ex. 6.13. Compliance: a randomized trial in which not all subjects complete the treatment course offered. Which groups should be compared?

One simple answer is to compare patients who received the new therapy with those who did not. The only group who received the new therapy is group N_1—those who were allocated the new treatment and completed it. Therefore we could compare group N_1 with either all other subjects ($C_1 + C_2 + N_2$), with all subjects allocated to the comparison treatment ($C_1 + C_2$), or with those allocated to the comparison treatment who completed it (group C_1).

However, if any of these comparisons is used, the value of randomization in controlling confounding is lost because the comparisons are no longer being made between randomized groups. There may be very considerable differences between the subjects who complete the new treatment (N_1), and those who withdraw or do not receive it (N_2). A classic example of this is shown in a randomized double-blind trial comparing lipid-reducing drugs with placebo, which was carried out in the USA between 1966 and 1969 [11] (**Ex.** 6.14). The outcome was mortality from any cause over the following 5 years. The mortality rate in those patients who were allocated to the lipid-lowering agent clofibrate, and actually consumed over 80 per cent of the allocated dosages, was 15.0 per cent. This rate could be compared with the mortality rate in all patients allocated to the
p. 178 placebo, which was 19.4 per cent, showing a statistically ↳ significant difference in favour of clofibrate. However, the patients who were randomized to clofibrate but did not take it had a much higher mortality (24.6 per cent). Of course, this is consistent with a beneficial effect of the drug; we would expect patients who did not

take the drug adequately to have a higher mortality than those who took it well. However, it would be dangerous to ascribe these results to the pharmacological effects of the drug. The point is dramatically illustrated by the mortality experience in relation to compliance with the placebo; those who took more than 80 per cent of the allocated doses of the placebo had a mortality of 15.1 per cent, and those who took less had a mortality rate of 28.2 per cent. The correct analysis of the randomized groups shows that the mortality rate in all those randomized to clofibrate was 18.2 per cent, similar to the rate in all those randomized to placebo (19.4 per cent). Any other comparison gives an incorrect result because of the confounding effects of factors related to compliance. The results also show that irrespective of what drug is prescribed, subjects who follow the instructions carefully have a lower mortality rate than those who do not. This cannot be ascribed to any pharmacological action, but reflects the influence of factors that are related to compliance, which are
p. 179 confounding factors in the association between the ↳ drug allocated and outcome. The correct analysis is referred to as *intention to treat* analysis; we compare the groups defined by the initial treatment offered.



Ex. 6.14. Compliance in a randomized trial: the results show the mortality rates from all causes after 5 years follow-up in patients randomized to either a lipid-lowering agent (clofibrate) or a placebo, in terms of their compliance with the drug regime. 'Compliant' means taking > 80 per cent of the allocated drug. From Coronary Drug Project Research Group [11]

Management and explanatory trials

The trial shown in Ex. 6.14 shows the dangers of making any other comparison in a randomized trial except between the groups chosen by randomization. This is despite the fact that, in many trials, the proportions of patients who fulfil all the protocol criteria, particularly that of maintaining the allocated therapy for its full course, may be relatively small.

This issue is important in both clinical and community-based studies. In the first randomized trial of breast cancer screening, women who were members of a health maintenance plan in New York were randomly allocated into two groups [12]. One group continued their normal medical care. Women in the other group were offered an innovative programme of screening for breast cancer, using mammography and clinical examination. About two-thirds of those offered this programme participated. The relevant outcome is death from breast cancer, and so the appropriate randomized analysis shows the effect of *offering* a breast cancer screening programme, with a two-thirds acceptance rate, on breast cancer mortality. After some 10 years of follow-up, there was a reduction of around 30 per cent in breast cancer mortality. This randomized trial shows the effect of offering a programme of screening. It does not measure the reduction in mortality risk for an individual who agrees to participate, except that this must be higher than the difference shown. Estimates of this can be made, but are uncertain because they depend on how those who participated in the programme

differed from those who declined the offered screening with regard to their underlying risk of death from breast cancer.

One could argue that analysis by comparison of randomized groups, while effective in controlling confounding, is likely to be inefficient, as there is a considerable dilution effect because many subjects in the intervention group are not actually receiving the intervention. However, in considering this we have to go back to the question of what the trial is actually for, and how the results will be applied. It is a naive supposition that all patients will follow a doctor's advice, or that everyone will participate in a planned intervention. In many situations, unless the intervention results in an improved outcome, given that a substantial proportion of subjects will not accept it in full, it will not be useful. We can distinguish between trials which address the *management* question—what is the effect of prescribing a certain therapy, or offering an intervention, in a practical situation?—and those whose aim is *explanatory*—irrespective of the practical issues involved, what is
p. 180 the effect of the intervention ↳ in ideal circumstances, in subjects who do accept it? This distinction between management and explanatory trials is important in the design and interpretation of randomized trials.

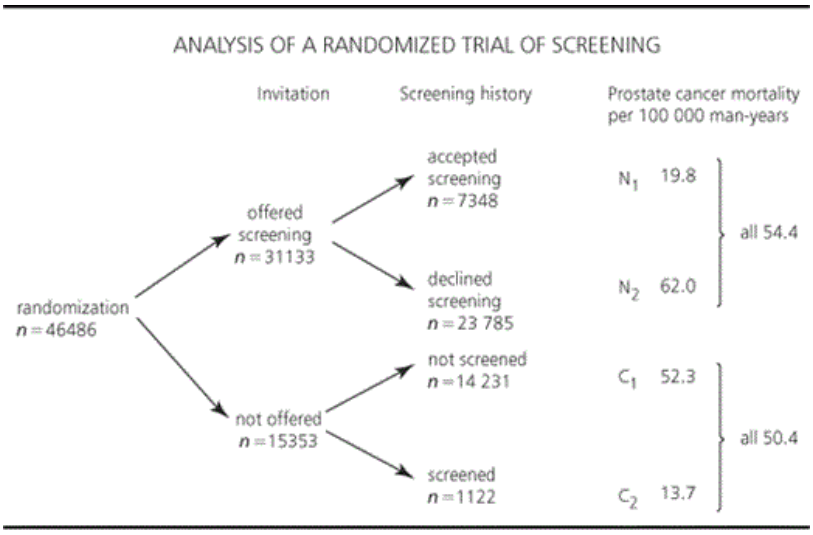
Both types of trial are legitimate; the dangers arise if the results from a study designed or analysed as an explanatory trial are interpreted in management terms. For example, iron deficiency anaemia in adolescence may affect cognitive function, and so iron supplementation could protect or improve cognitive function. A randomized trial of iron supplementation carried out on high school girls in the USA assessed this, and showed that, after 8 weeks, the girls who received iron supplementation performed better on tests of verbal learning and memory than girls in the control group [13]. However, girls who were clinically anaemic were excluded from the trial, other vitamins or iron supplements were prohibited, many strategies were used to ensure compliance, and the outcome measures used were research tools; most important, the analysis was based on only the girls who completed treatment rather than being an intention to treat analysis. As pointed out in an accompanying editorial, this is an explanatory study which gives the results under ideal conditions, and therefore demonstrates the potential rather than the practical value of iron supplementation [14]. A management study, to assess directly whether iron supplementation would have clinically relevant benefits in practice, would need to use realistic rather than ideal methods of ensuring compliance, less stringent eligibility criteria, and a more relevant endpoint and time of assessment, and in particular include in the analysis all subjects randomized.

Randomized trials are often designed to guide clinical or health policy decisions, and the terms *pragmatic* or *practical* clinical trials have been used for trials designed to facilitate such decision-making. These have some features that contrast with more explanatory orientated trials, which may concentrate on a highly selected group of subjects. Practical trials relevant to policy-making should be based on relevant interventions, and include a diverse population of study participants, recruited from different practice settings, with data collected on a broad range of health outcomes [15].

If a management trial, correctly analysed by intention to treat, shows no effect, this could be because the intervention did not work or because too few subjects accepted it. A controversial example, which also shows the dangers of incorrect analysis, is a randomized trial of prostate cancer screening. Men aged 45–80 years sampled from the electoral rolls in Quebec City, Canada, were randomly allocated in a 2:1 ratio either to receive an invitation to a screening programme using annual measurements of serum prostate-specific antigen (PSA)
p. 181 and digital rectal examination, or to be in a non-invited control group. ↳ This is a standard randomized design. The outcome was the death rate from prostate cancer in the 11 years after randomization, reported in two main papers [16,17].

The results based on this randomization (**Ex.** 6.15) show a relative risk of 1.08, i.e. an 8 per cent increase in deaths in the group offered screening. However, in contrast, the summary of the more recent paper reports a relative risk of 0.38, i.e. a 62 per cent reduction in mortality. The earlier paper showed a similar reduction, and the authors concluded that their study 'demonstrates, for the first time, that early diagnosis and treatment

permits a dramatic decrease in deaths from prostate cancer' [16]. The large mortality reduction is based on a comparison between men who were randomized to screening and were screened, and those randomized to no screening who were not screened (groups N_1 and C_1 in the terminology of Ex. 6.13, shown earlier). This analysis restricted to compliant subjects is shown in Ex. 6.15, and gives the relative risk of 0.38. In the earlier paper, the main comparison was between all the men who were screened, whether they were invited for screening or not, and all other men in the study who were not screened (groups $N_1 + C_2$ versus groups $N_2 + C_1$ in Ex. 6.13); this gave a similar large reduction.



Ex. 6.15. Different analyses of a randomized trial of screening: the correct analysis, maintaining the randomized groups, gives a relative risk of 54.4/50.4 = 1.08. Other comparisons lose the advantage of the randomization. From Labrie *et al*. [17]

p. 182 In both these analyses, the benefits of the randomized design have been lost. In the body of the recent paper an analysis on an 'intent to screen' basis was also made and the lack of effect recognized, but this is given little importance. In this trial, few men invited for screening were actually screened (23 per cent), so that the randomized trial has lost most of its ability to detect an effect of screening because of misclassification (in addition, 7 per cent of the men randomized to the non-invited group were screened, but this is a smaller problem). However, this does not give validity to the non-randomized comparisons which, presented in papers with the title of a randomized trial, are misleading. It is surprising that such a fundamentally flawed analysis was published in a reputable and peer-reviewed journal on a topic of great practical importance. The problems in the analysis have been pointed out in correspondence [18,19].

Stratification

To discuss the next method of controlling confounding, let us go back to the example of studies of oral contraceptives and myocardial infarction, and the problem of confounding by smoking. As we have seen, we could use restriction to avoid confounding by smoking. This could lead to a series of restricted studies, one looking at non-smokers, one at smokers of a certain quantity, and so on. However, this situation can be achieved in one study. We could include oral contraceptive users and non-users in the study without any limitations, but with careful recording of each woman's smoking history. Then in the analysis, we divide the data, and compare oral contraceptive users who have never smoked with non-users who have never smoked, and so on throughout the various levels of smoking. This procedure is *stratification*, as strata or layers of data are formed. The stratification is done on the basis of the suspected confounding factor. This is the most important and widely used method of controlling confounding. It is used in the majority of studies either on its own or in combination with other methods.

Indeed, if you have worked through this chapter, you have already used stratification several times. In Ex. 6.4 we demonstrated the effect of stratification by obesity in a study relating exercise to myocardial infarction. In Ex. 6.6, we stratified by the size of the renal stone in the study comparing surgical methods in the treatment of renal stones. In Ex. 6.7, we used four age groups to control confounding by age in a cohort study. In Ex. 6.8, we used stratification by hair colour in a case–control study relating eye colour to skin cancer. In each of these examples, we went as far as dividing the data into strata, and looking at the measure of association (relative
p. 183 risk or odds ratio) within each stratum of the confounder, and we also made some comments about the range ↳ of these results. The extra process, which we will now consider, is how to use the stratified data to produce a *single overall estimate* of effect that has been *adjusted* for the effects of the confounding factor.

Stratification in cohort and intervention studies: Mantel–Haenszel method

The results from the cohort study relating exercise to deaths from heart disease previously shown in Ex. 6.7 are
p. 184 summarized in **Ex.** 6.16, Subtable A, ↳ for the four age groups. The summed data for all ages give a relative risk of 3.44, but this is misleading as it is confounded by age; the age-specific relative risks range from 1.06 to 1.99.

STRATIFICATION IN A COHORT STUDY: MANTEL–HAENSZEL ANALYSIS

Subtable A: Data stratified by the confounder (age)

Age group	*Exposed (light exercise)*			*Unexposed (heavy exercise)*			Relative Risk
	Deaths	Man-years (thousands)	Rate/1000	Deaths	Man-years (thousands)	Rate/1000	
	a_i	N_{1i}	a_i/N_{1i}	c_i	N_{0i}	c_i/N_{0i}	
35–44	3	5.9	0.51	4	8.3	0.48	1.06
45–54	62	17.6	3.52	20	11.0	1.82	1.94
55–64	183	23.7	7.72	34	7.4	4.59	1.68
60–74	284	17.8	15.96	8	1.0	8.00	1.99
All ages	532	65	8.18	66	27.7	2.38	

Crude relative risk (no adjustment for age) = 8.18/2.38 = 3.44

Subtable B. Calculation of summary relative risk, adjusted for the confounder, by the Mantel–Haenszel method

Age group	Total man-years (T_i)	a_iN_{0i}/T_i	c_iN_{1i}/T_i
35–44	14.2	1.75	1.66
45–54	28.6	23.85	12.31
55–64	31.1	43.54	25.91
65–74	18.8	15.11	7.57
Sum		84.25	47.45

Mantel–Haenszel relative risk = (84.25/47.45) = 1.78

Ex. 6.16. Stratification in a cohort study: the study shown in Ex. 6.7 gives, on simple analysis, a misleading result because of confounding by age (subtable A). Stratification by age avoids this. One method of calculating a summary relative risk measure, adjusting for the stratification, is shown in subtable B

A practical difficulty arises when data are divided into a number of strata. The numbers of subjects in one stratum may be small, and therefore the outcome measure, such as the relative risk, for that stratum will be imprecise; therefore the relative risk estimates may vary considerably among the strata. There are two issues. First, can we assume that there is a *constant* measure of association, such as relative risk, which can be applied to all strata? At this point we will simply point out that differences based on small numbers are unlikely to be important (such as the youngest age group in Ex. 6.16), and we should be concerned only about substantial differences between the relative risks in different strata. In the next chapter we will present statistical methods to test whether the assumption of a constant relative risk for all the strata is justified. Then, if we assume that the true relative risk is likely to be the same in all strata, we can use the data for all strata to produce a single estimate of relative risk, which is adjusted for the effects of the confounding factor.

This adjusted relative risk will be some type of average of the relative risks in the different strata. There are several methods of producing an overall estimate, which use different methods of averaging. Indeed, one method would be a simple arithmetic average of the relative risks for each stratum. However, this would assume that they are equally precise, and in our example the relative risk for the youngest age group is based only on seven deaths in total, and intuitively we should give it less importance than the relative risks in the other groups.

A more appropriate method of averaging is to give each stratum an importance related to the amount of information or, in other words, the numbers in the stratum. One widely used method is shown in **Ex.** 6.17, and applied in Ex. 6.16 in Subtable B. The formula for the adjusted relative risk has the same structure as the

formula for the relative risk in one stratum. The numerator is the sum of the numerator terms derived from each stratum, and the denominator is a sum of the denominator terms derived from each stratum. The result gives a *weighted average* of the relative risks, in which the weights given to each stratum relate to the amount of information within each stratum. This adjusted relative risk is a reasonable summary value of the overall relative risk, adjusting for the confounding effect of the stratification factor, in this case, age. Here this adjusted relative risk is 1.78, which is compatible with the age-specific relative risks. This particular method is derived
p. 185 from a method first developed for use in case–control studies, and is referred to as the ↳ Mantel–Haenszel method [20]. (The algebraic notation used in Exs. 6.16 and 6.17 is the same as that used previously in Chapter 3, except that we indicate the data within one stratum by a subscript 1 for the first stratum, 2 for the second, etc., which is shown in general terms by the subscript *i*).

MANTEL–HAENSZEL ESTIMATION OF RELATIVE RISK

In each subtable *i*, stratified by the potential confounder:

	Cases	Total subjects or person-years
exposed	a_i	N_{1i}
unexposed	c_i	N_{0i}
		T_i = total in the subtable

Relative risk for this subtable = $(a_i/N_{1i})/(c_i/N_{0i}) = a_iN_{0i}/c_iN_{1i}$

The Mantel–Haenszel estimate of relative risk uses the data from all subtables, but gives an estimate of the unconfounded relative risk:

given by $$\frac{\sum_i a_iN_{0i}/T_i}{\sum_i c_iN_{1i}/T_i}$$

where $\sum_i$ indicates summation over all the subtables

Ex. 6.17. Stratification: calculation of the Mantel–Haenszel estimate of the summary relative risk in a stratified analysis

Stratification can be used in any study design, including randomized trials. We emphasized earlier that while randomization is likely to produce evenly balanced groups, it does not always do so. In Ex. 6.12, some information from a randomized trial of tolbutamide and placebo in the management of diabetes was given, which suggested some substantial differences in the groups being compared. **Exhibit** 6.18, Subtable A, gives the main results of that study, showing death rates from all causes of 14.7 per cent in the tolbutamide-treated patients and 10.2 per cent in the placebo-treated patients, giving a relative risk of 1.44. We can use stratified analyses to assess whether these differences are influenced by the differences in the baseline characteristics shown in Ex. 6.12. In Ex. 6.18, Subtable B, the stratification by arterial calcification is shown. Arterial calcification is a true confounder; it is more common in the tolbutamide-treated group than in the placebo-treated group, and within each treatment group the death rate is higher in the subjects with arterial calcification. However, both within subjects with arterial calcification and within those without arterial
p. 186 calcification, the death rate is higher in the tolbutamide-treated subjects. ↳ The best estimate of the relationship between tolbutamide treatment and mortality, adjusted for the differences in prevalence of arterial calcification, is given by the Mantel–Haenszel relative risk calculation, which gives a relative risk of 1.32. We compare this with the original relative risk of 1.44 in Subtable A. The conclusion is that the crude increased risk in the tolbutamide group (*RR* = 1.44) is reduced by stratifying for arterial calcification, but is not abolished, and most of the excess risk is still maintained (*RR* = 1.32). Analyses of this form can be carried out for each of the variables listed in Ex. 6.12. However, it is clear that this method of analysis is still unsatisfactory, as although it is easy to adjust for each of the variables singly, we are still not answering the more general question: does the
p. 187 difference in mortality between tolbutamide- and placebo-treated patients persist when we take into account

↳ the effect of *all* the factors on which we have baseline information? That issue requires more complex analysis which will be shown later in this chapter.

STRATIFIED ANALYSIS OF A RANDOMIZED TRIAL

Subtable A. Main results (n = 409)

Treatment	Deaths	Survivors	Total	% dead	Relative risk
Tolbutamide	30	174	204	14.7	1.44
Placebo	21	184	205	10.2	1.0(R)

Subtable B_1. Stratified: arterial calcification present (n = 68)

Treatment	Deaths	Survivors	Total	% dead	Relative risk
Tolbutamide	13	26	39	33.3	1.93
Placebo	5	24	29	17.2	

Subtable B_2. Stratified: arterial calcification absent (n = 333)

Treatment	Deaths	Survivors	Total	% dead	Relative risk
Tolbutamide	16	143	159	10.1	1.09
Placebo	16	158	174	9.2	

Relative risk adjusted for arterial calcification
$= (13 \times 29/68 + 16 \times 174/333)/(5 \times 39/68 + 16 \times 159/333) = 1.32$

Ex. 6.18. Stratified analysis of a randomized trial: main results of the trial shown in Ex. 6.12 (subtable A), and results stratified for arterial calcification (subtables B_1 and B_2), and adjusted relative risk calculated by the modified Mantel–Haenszel method (Ex. 6.17). From University Group Diabetes Program [10]. As often occurs, some data are missing: eight subjects are omitted from the stratified analysis as data on arterial calcification were missing. The crude relative risk for the two subtables combined is 1.42

Direct standardization

Another method of taking a weighted average of stratum-specific relative risks is direct standardization, which is frequently used in large data sets such as vital statistical data from whole countries or communities. It is most frequently applied to age, often within the 18 5-year age groups from 0–4 years to 85+ years, or to sex, or to ethnic group. It is simply another form of averaging of stratified data, and can be applied to any confounder. For each group being compared, the stratum-specific rates are multiplied by the number of subjects or person-years in that stratum of a 'standard population'. This standard population may be an intuitively relevant population (such as the whole country in a particular year), or an arbitrary or even artificial population.

Exhibit 6.19 shows the use of direct standardization to compare incidence rates for two populations, taking account of the differences in age distribution. The data show the incidence rates of cancer of the stomach in men recorded between 1993 and 1997 by cancer registries in India (Mumbai) and Sweden, showing total rates and age-specific rates [21]. The crude incidence rates are 3.6 per 100 000 person-years in India and 17.2 per 100 000 person-years in Sweden, giving an incidence ratio of 17.2/3.6 = 4.8, i.e. the crude incidence rate is nearly five times higher in Sweden than in India. To adjust for age, we will use the 'world standard population', an arbitrary population developed by the World Health Organization for this purpose which approximates the age distribution of the whole world population. For each age group the observed rate is multiplied by the world standard population number for that age group, and then the results are summed and divided by the total world standard population. The direct age-standardized incidence rates produced in this way are 6.4 per 100 000 person-years in India and 8.6 per 100 000 person-years in Sweden. The ratio of these is only 1.35, showing that the incidence rates, once adjusted for the differing aged populations, are only modestly higher in Sweden than in India. Stomach cancer, like most cancers, increases greatly in incidence with increasing age. The Indian

population is younger than the world standard population, and so its crude incidence rate is low and its age-standardized incidence rate is higher than the crude rate. The Swedish population is considerably older than the world standard population, and so its crude rate is high and adjustment produces an age-standardized rate which is lower than the crude rate. Most of the difference in crude rates between the two countries is due to the older age distribution in Sweden, rather than to the

60–64	146	624 025	23.4	4	93.6	291	984 656	29.6	4	118.2
65–69	161	372 010	43.3	3	129.8	480	947 266	50.7	3	152.0
70–74	114	225 650	50.5	2	101.0	712	907 764	78.4	2	156.9
75+	129	240 825	53.6	2	107.1	1756	1 405 204	125.0	2	249.9
Totals	1063	29 561 285		100	635.2	3750	21 746 247		100	860.0
Crude incidence rate			= 1063/295.61285		3.60			=3750/217.46247		17.24
Age-standardized incidence rate			= 635.2/100		6.35			=860/100		8.60
Incidence rate ratio (Sweden:India)										
Crude			= 17.24/3.60	4.80						
Age-standardized			= 8.60/6.35	1.35						

Ex. 6.19. Age standardization of incidence rates: data for two populations (Mumbai and Sweden), adjusted by direct standardization using the World Health Organization 'world standard' arbitary population. Data from *Cancer Incidence In Five Continents*, Vol. VIII, [21]

p. 188

AGE STANDARDIZATION

Cancer of the stomach in men; data from cancer registries, 1993–1997

	India (Mumbai)					Sweden				
Age	Cases	Person-years	Incidence per 100 000 person-years	World standard population	Rate*std population	Cases	person-years	Incidence per 100 000 Person-years	World standard population	Rate*std population
0–4	0	2 606 530	0.0	12	0.0	0	1 491 716	0.0	12	0.0
5–9	0	2 873 805	0.0	10	0.0	0	1 454 119	0.0	10	0.0
10–14	0	2 814 265	0.0	9	0.0	0	1 290 659	0.0	9	0.0
15–19	2	2 830 345	0.1	9	0.6	1	1 313 792	0.1	9	0.7
20–24	4	3 351 470	0.1	8	1.0	1	1 465 616	0.1	8	0.5
25–29	13	3 092 025	0.4	8	3.4	3	1 607 164	0.2	8	1.5
30–34	18	2 678 690	0.7	6	4.0	16	1 591 005	1.0	6	6.0
35–39	36	2 388 920	1.5	6	9.0	20	1 497 476	1.3	6	8.0
40–44	61	1 918 280	3.2	6	19.1	46	1 515 350	3.0	6	18.2
45–49	89	1 532 425	5.8	6	34.8	83	1 653 006	5.0	6	30.1
50–54	145	1 194 715	12.1	5	60.7	149	1 488 244	10.0	5	50.1
55–59	145	817 305	17.7	4	71.0	192	1 133 210	16.9	4	67.8

p. 190 modest difference in the incidence rates of stomach cancer within age groups. Of course, other factors, such as the completeness of diagnosis and recording, also have to be considered.

Direct standardization can be applied to any stratified data where the stratum-specific rates are known. **Exhibit** 6.20 shows direct age standardization applied to the study of exercise and heart disease shown earlier. The death rates are calculated within each age group and then, for both the light and the heavy exercise group, these rates are multiplied by the standard population, which here is taken as the total number of man-years of observation in each age group. The product of the actual death rate and the standard population number gives an 'expected' number of deaths in each stratum. For each exercise group, the total of these expected numbers is the number of deaths which would have occurred if the distribution by age had been the same as that of the standard population. As the same standard population has been applied to each of the exercise groups,

p. 191 comparison of these expected numbers removes ↳ the confounding effect of age (unless there is so much variation within the age strata used that there is still some confounding, in which case narrower age strata should be used). The ratio of these expected numbers gives the age-standardized relative risk, which here is 1.84. This is not numerically the same as the Mantel–Haenszel relative risk, which as shown in Ex. 6.16 was 1.78, because a different weighting system has been used. In the example given, the Mantel–Haenszel estimate is probably preferable, as its weighting system takes into account the different amount of information in each stratum, whereas the age-standardization system does not necessarily do this. In large data sets, such as routine vital statistics, there is so much data available that this is not really a problem, and that is where direct standardization is most frequently used.

STRATIFICATION IN A COHORT STUDY: DIRECT STANDARDIZATION

Age group	Standard population	*Exposed (light exercise)* Observed death rate	Expected deaths in std. pop.	*Unexposed (heavy exercise)* Observed death rate	Expected deaths in std. pop.
	S_i	r_{ei}	$S_i r_{ei}$	r_{ui}	$S_i r_{ui}$
35–44	14.2	0.51	7.2	0.48	6.8
45–54	28.6	3.52	100.7	1.82	52.1
55–64	31.1	7.72	240.1	4.59	142.7
65–74	18.8	15.96	300.0	8.00	150.4
Sum	92.7		648.1		352.0

Age-standardized incidence rate =
(sum expected deaths) / (sum standard population) = 6.99 (exposed); 3.80 (unexposed)
Age standardized relative risk = 6.99/3.80 = 1.84

Ex. 6.20. Stratification in a cohort study: direct standardization. The stratified data are the same as in Ex. 6.16. Here a standard population is chosen as the total man-years at risk in each age group, and to this is applied the age-specific death rates in each of the exposure groups, to give the numbers of deaths expected in each exposure group if each had had the same age distribution. The ratio of these expected numbers is the standardized relative risk

Indirect standardization

Another method of standardization is indirect standardization, which gives results often referred to as a standardized mortality ratio (SMR) or standardized incidence ratio (SIR). This method is frequently used to compare one smaller group with a larger group that includes it, for example to compare rates in one occupational group with those of the whole country. **Exhibit** 6.21 shows an example where 12 deaths from lung cancer occurred in an occupational group (male cooks) over 4 years in New Zealand. To assess this, we calculate the 'expected' number of lung cancers, i.e. the number that would arise if the death rates for the standard population (here, all employed men) applied to the cooks. For each age group, we multiply the age-specific rate in the standard population by the numbers of person-years in each age group for the cooks (basically the number of cooks, obtained from census data, multiplied by 4 years) to give an expected number of deaths in that age group. We then add the expected numbers, here giving 7.1 'expected' deaths. The ratio of the observed number of deaths to this expected number of deaths shows how the actual mortality experience varies from that expected, adjusted for age. This is often expressed as a percentage, so that a value above 100 shows a higher mortality rate in the study population than in the standard population. In our example, the (age) standardized mortality rate (SMR) is 1.69, or 169 per cent. For this calculation, the only required data for the cooks is the number of cooks (or person-years) in each age group and the total number of deaths. It is not necessary to know the number of deaths within each age group, which is needed for direct standardization or a Mantel–Haenszel analysis. However, this indirect standardization method is not so useful, as although it produces a comparison of each specific population with the reference standard population, it is not so simple to
p. 192 compare ↳ the SMR of one specific group with that of another group. To make a comparison between two specific groups, it is better to use direct standardization or the Mantel–Haenszel method.

CALCULATION OF A STANDARDIZED MORTALITY RATIO

Total observed number of events (deaths from lung cancer) = 12

Expected number of deaths if stratum-specific rates were same as the standard population is calculated as follows:

Strata of confounder (age group)	Number of person-years P_i	Mortality rate in standard population per 100 000 R_i	Expected number of deaths = $R_i \times P_i$
25–34	7557	0.8	0.06
35–44	4023	4.8	0.19
45–54	2961	47.9	1.42
55–64	2043	265.8	5.43

Total expected number = sum of expected numbers = 7.10

Standardized mortality ratio = observed / expected = 1.69

Ex. 6.21. Calculation of a standardized mortality ratio: the data needed are the total number of events in the 'special' population. Here the special population is male 'cooks and related workers' aged 25–64 in New Zealand, 1983–1986. The expected number is calculated by applying the age-specific rates for a 'standard' population, here all employed males in New Zealand, 1983–1986, to the age-specific population at risk for the special occupational group

Stratification in case–control studies

To apply stratified analysis in a case–control study, we follow the same principle: we divide the data into strata defined by levels of the confounding factor, and calculate the measure of association, usually the odds ratio, for each stratum. As the same issues of small numbers and consequent instability of estimates apply in case–control studies as in cohort studies, we need a method of producing a summary measure of association. The most widely used method is the Mantel–Haenszel estimate of odds ratio, described in a now classic paper [20]. This is a weighted average of the stratum-specific odds ratios, with the weights being dependent on the numbers of observations in each stratum, as shown in **Ex.** 6.22.

MANTEL–HAENSZEL ESTIMATION OF ODDS RATIO

In each subtable i, stratified by the potential confounder:

	Cases	Controls
exposed	a_i	b_i
unexposed	c_i	d_i

T_i = total in the subtable

Odds ratio for this subtable = $a_i d_i / b_i c_i$

The Mantel–Haenszel estimate of odds ratio uses the data from all subtables, but gives an estimate of the unconfounded odds ratio:

given by $\sum_i (a_i d_i / T_i) \Big/ \sum_i (b_i c_i / T_i)$

where $\sum_i$ indicates summation over all the subtables

Ex. 6.22. Stratification: calculation of the Mantel–Haenszel estimate of the summary odds ratio in a stratified analysis

p. 193 **Exhibit** 6.23 shows some simplified, but real, data from a case–control study comparing patients with malignant melanoma (a skin cancer) with community-based controls [22]. Table A shows a positive association of melanoma with a history of severe sunburn, giving an odds ratio of 1.40. This association could

be due to the trauma of the sunburn, the sun exposure involved, or the individual's susceptibility to sunburn. In the subtables, we control for susceptibility to sunburn. Both in subjects who burn easily and in those who do not, the odds ratio for the association between sunburn history and melanoma is lower than that in the crude analysis; the Mantel–Haenszel estimate gives 1.19 as the unconfounded odds ratio. Thus the crude odds ratio of 1.40 is produced partially by confounding by tendency to sunburn, which is also related to melanoma. Statistical tests on these data will be shown in Chapter 7 (Ex. 7.8).

Effect modification

In the example in Ex. 6.23, the odds ratios in individual strata are so similar that there is no obvious utility in the summary estimate. However, with many subtables with small numbers of observations in each, the stratum-specific odds ratios will be unstable, but the summary estimate provides a stable measure of the overall odds ratio which is not affected by confounding by the factor which has been stratified. Of course, if the
p. 194 odds ratios in the different ↳ strata are very different from each other, it may be misleading to use a summary estimate.

In a case–control study relating smoking to carcinoma of the uterine cervix, the overall data show an odds ratio of 6.6 (**Ex.** 6.24). When the data are subdivided by age, the odds ratios show great variation, from 27.9 at age 20–29, to 5.9 at age 30–39, and 2.8 at ages over 40 [23]. This large variation in the association between cervical cancer and smoking demonstrates *effect modification* (also called *interaction*). Effect modification can occur
p. 195 without confounding. To assess confounding by age, we can use a Mantel–Haenszel estimate, calculated ↳ by the method already shown in Ex. 6.22, which gives an odds ratio after age adjustment of 6.3. This is close to the crude odds ratio of 6.6, and shows that there is very little confounding by age in these data.

This image cannot be displayed online for copyright reasons.

This example illustrates the difference between confounding and effect modification. If a factor is acting solely as a confounding factor, it will bias the overall association between exposure and outcome in the data set, and this can be accomplished without any variation in the measure of association between different strata. The examples given so far demonstrate this. In Ex. 6.23, the odds ratio between melanoma and sunburn history was confounded by tendency to sunburn, but there was no substantial effect modification, i.e. the association was virtually the same amongst subjects who sunburnt easily and those who did not. In the study of exercise levels and heart disease in Ex. 6.20, there was also substantial confounding by age, but little evidence of any effect modification, i.e. the relative risks comparing light and heavy exercise workers were similar in the different age groups. In contrast, in the current example in Ex. 6.24, there is relatively little confounding, but there is very clear effect modification. A similar result was seen in the randomized trial analysis in Ex. 6.18 where the
p. 196 relative risk between tolbutamide treatment and mortality was ↳ 1.44 in the crude form, and 1.32 after adjustment for arterial calcification, showing only a small degree of confounding. However, the relative risk for subjects with arterial calcification (RR = 1.93) was substantially higher than for the subjects without such calcification (RR = 1.09), showing effect modification. We shall explore the example of cervical cancer and smoking further in Chapter 7, Ex. 7.9.

Exhibit 6.25 shows results from a cohort analysis of the case fatality rate from SARS (severe acute respiratory
p. 197 syndrome) in Hong Kong [24]. The overall ↳ fatality rate was considerably greater in men than in women, with a relative risk of 1.66. Stratification by age demonstrates both confounding and effect modification. As shown in Ex. 6.25, some of the difference in mortality rates is due to the older age distribution of the men, allied to the increasing death rate with increasing age. This is shown by the Mantel–Haenszel relative risk, adjusted for age, being 1.35, i.e. lower than the crude relative risk. Examination of the results within age groups shows

that there is also effect modification, with the higher death rate in men being more pronounced in the younger age groups, and there being no substantial difference between the death rates in men and women at ages 75 years and over.

CONFOUNDING AND EFFECT MODIFICATION

Fatality rate in SARS (severe acute respiratory syndrome), comparing males with females, in Hong Kong

Table A: all ages

		cases	deaths	survivors	fatality rate (%)	male: female relative risk
	Males	776	170	606	21.9	1.66
	Females	979	129	850	13.2	
	Total	1755	299	1456	17.0	

Subtable B: data stratified by age

		cases	deaths	survivors	fatality rate (%)	male: female relative risk
age 0–44	Males	425	27	398	6.4	2.27
	Females	607	17	590	2.8	
age 45–74	Males	249	77	172	30.9	1.45
	Females	295	63	232	21.4	
age ≥ 75	Males	102	66	36	64.7	1.02
	Females	77	49	28	63.6	

Mantel–Haenszel relative risk, adjusted for age = 1.35

Ex. 6.25. Combined confounding and effect modification: comparison of male and female case fatality rates from SARS in Hong Kong. The death rates are higher in males, particularly at younger ages (effect modification by age), and the overall crude relative risk of 1.66 is partially due to confounding by age, shown by the adjusted relative risk being only 1.35. From Karlberg *et al*. [24]

Thus in an exposure–outcome relationship, a third factor may act as a confounding factor or an effect modifier, may have both effects, or may have neither effect.

Matching

Another method of avoiding confounding is to choose the comparison subjects for a study so that they are similar to the case or exposed subjects with regard to specified confounding factors, and then analyse the study appropriately. For example, in a cohort study assessing the risk of myocardial infarction in women who use oral contraceptives, factors such as age and cigarette smoking might be important confounders. If, for each woman using oral contraceptives entering the study, we select a comparison woman who is not using oral contraceptives but is the same age and has the same smoking history, we will create two groups of subjects who are similar in terms of these two factors. Therefore within our study there will be no association between age or smoking status and oral contraceptive exposure, and thus these two factors will not be confounding. This process is referred to as matching.

Matching is a much more complex technique than it appears, and the applications and value of the method are very commonly misunderstood. This misunderstanding often arises from a lack of appreciation of the other methods of controlling confounding, particularly stratification, so that the specific advantages and disadvantages of matching are not recognized. More subtly, difficulties arise because matching has three different purposes: it can be used to increase the *efficiency* of the study, to *control confounding*, or to improve the *comparability* of the information collected.

Frequency matching to increase efficiency

With regard to *efficiency*, i.e. the amount of information that is gained in relation to the size of the study, the
p. 198 value of matching is quite simple. In the ↳ example here, it is likely that most women using oral contraceptives will be between 16 and 45 years old. There is no value in enrolling as comparison subjects women who are 60 years old or, even more obviously, enrolling men in the study. Ideally the ratio of comparison subjects to exposed subjects in each age range should be reasonably constant. An age group in which there are many oral contraceptive users and very few non-users, or vice versa, will give unreliable information because of the small numbers in one category (in the extreme, a group with no exposed or no controls gives no information). The same process can be used for other factors such as smoking history. Thus to increase efficiency, matching can ensure that the groups of subjects being compared are similar in terms of the most important potential confounding factors in the study. This can be achieved by 'frequency matching'; the study is designed so that the distribution of the groups of subjects being compared is similar in terms of major confounding factors. This is an easily applied, useful, and widely used technique, particularly for such characteristics as age and sex. Frequency matching should be considered only as a method of ensuring reasonable efficiency, and should not be regarded as a method of controlling confounding, because it involves only approximate matching and does not ensure total comparability. The confounding factors, both those on which frequency matching has been done and others, need to be controlled by methods such as stratification or multivariate analysis. The analysis is done in the same way as in unmatched studies. Indeed, the advantage of frequency matching is to increase the efficiency of stratification and multivariate methods to control confounding.

Individual matching to control confounding

The second form of matching, 'individual matching', is a precise technique. Rather than simply ensuring broad comparability between the groups being compared, comparison subjects are chosen to match particular index subjects with regard to one or more specified confounding factors. The purpose is not only to improve efficiency, but also to control for the confounding effects of these factors.

There is an important distinction between the use of this technique in cohort or intervention studies and in case–control studies. In cohort and intervention studies, precise matching by important confounding factors will control for the effects of those factors. An analysis done in the same way as in an unmatched cohort study will give estimates of relative and attributable risks which can normally be regarded as free from the confounding effects of the variables on which matching has been done. A caution applies, as during the time course of the study subjects may withdraw or be lost to follow-up, and so the initial matched situation will be
p. 199 compromised. This can be dealt ↳ with by stratifying or using multivariate methods to deal with any confounding, or in some studies it may be necessary to censor the whole matched set of subjects at the same time, that is, to terminate their follow-up (see Chapter 7).

In a case–control study, controls are selected by matching them to the cases on specified factors. This will give an efficient study, but to control confounding it is necessary to analyse the study by special methods that keep the data in matched form, i.e. consider each matched group of cases and controls as a unit. As long as this is done correctly, the combination of matching and using a matched analysis should control confounding by the

matching factors. The analysis is basically a stratified analysis in which each matched case–control set forms one stratum, or the equivalent multivariate analysis. This is necessary because matching can in fact introduce confounding by the matching factors, and this confounding needs to be controlled. If a matched case–control study is not analysed correctly, the result may be invalid.

The simplest form of analysis is given by a fixed one-to-one matching ratio, where one matched control is
p. 200 chosen for each case, as shown in **Ex.** 6.26 [25]. ↳ Note the format of the table: it shows the *numbers of pairs* classified by exposure and by outcome. The odds ratio is simply the ratio of the number of discordant pairs where only the case is exposed to the number of discordant pairs where only the control is exposed. On the null hypothesis, these numbers will be equal. The (incorrect) unmatched table in the familiar 2 × 2 format uses exactly the same data, but the value of the matching is lost. If the matching factors, as in this example, have a strong confounding effect, the odds ratio based on an unmatched analysis is confounded and is substantially different from the matched unconfounded odds ratio. It will be biased towards the null, as shown here. Where another fixed ratio of controls to cases (such as 3:1) is used, the analysis is also relatively simple, but where the matching ratio is variable it becomes complex. Analyses of a fixed-ratio matched study are shown in Appendix Tables 5 and 6. Matched analyses also become complex if it is necessary to adjust for further confounding factors after the matched design is set up. Normal stratification procedures are difficult to use, as the matched groups must be kept intact. However, 'conditional' multivariate statistical techniques will allow a matched design to be handled in an analysis that controls other confounding factors; this will be discussed later in this chapter.

ONE-TO-ONE MATCHING IN A CASE–CONTROL STUDY

A. Distribution of 120 case-control pairs by smoking history

		Controls	
		Smokers	Non-smokers
Cases	smokers	31	30
	non-smokers	7	52

Matched odds ratio = 30/7 = 4.3

B. An incorrect unmatched analysis of the same 240 individuals

	Cases	Controls
smokers	61	38
non-smokers	59	82
	120	120

Unmatched odds ratio = (61 × 82)/(59 × 38) = 2.2

Ex. 6.26. One-to-one matching in a case–control study: this yields a simple analysis. Here 120 male patients with nasal or nasal sinus cancer (a very rare cancer), seen in one clinic over a 38-year period, were matched to male controls with a range of other non-smoking related cancers by age and year of diagnosis. Table A shows a matched analysis; the odds ratio is based only on the pairs with different exposures. Table B shows the unmatched analysis of the same data; this analysis is incorrect as it loses the value of the matched pair comparison, and produces a substantially different odds ratio. From Elwood [25]

Matching to increase comparability of data

A further use of matching is to ensure comparability in terms of the information collected. In a case–control study where some case subjects are interviewed in a hospital and others in their own homes, it is logical to ensure that the comparison subjects are matched for method of interview, and ideally the interviews are conducted on a single-blind basis. This was done in a case–control study of venous embolism and hormone replacement therapy, where most interviews took place in hospital, but 23 per cent were done after discharge [26]. Similar considerations arise if several investigators or centres are involved. However, stratification or other methods may also be used.

Advantages of matching

What then are the advantages of matching? There are three main situations in which matching may be useful, but these advantages always have to be compared to the disadvantages of matching when designing a study.

First, matching is useful where there is a complex confounding factor. Matching has particular value where
p. 201 there is an important confounding variable that cannot be easily measured or easily defined. Examples include
↳ complex social factors, medical care factors, environmental exposures, or circumstances in childhood; these might be controlled by matching with neighbours, other patients, coworkers, or siblings, respectively. For example, to assess associations with emotional and behavioural problems, girls with anorexia nervosa were compared with their unaffected siblings, giving control for a range of family and childhood factors [27]. Similarly, to assess the relationship of smoking with Parkinson's disease, using sibling controls may control for other environmental exposures [28].

Studies comparing individuals who are twins with their co-twins, where they differ on the exposure or outcome of interest, can be a very powerful design. Thus an international case–control study comparing women with breast cancer at ages under 50 with their unaffected twin sisters showed associations with childhood weight, height, and time of breast development [29], and in a prospective cohort study twins with moderate alcohol consumption (compared with those with lower consumption) showed a reduced incidence of type 2 diabetes than their same-sex co-twins, despite the similarity in genetic and early life factors [30]. Comparison between monozygotic (identical) and dizygotic (non-identical) twins is a fundamental technique in assessing the contribution of genetic and environmental factors. Using a twin registry in Sweden, smokers were compared with their non-smoking co-twins; from 9319 pairs of twins, 1924 pairs who differed in their smoking habits were found; the prevalence of cough and bronchitis was higher in the smoking twins, and the prevalence ratio was similar for non-identical and for identical twin pairs, showing that the association with smoking could not be explained by a genetic confounding factor [31].

Secondly, matching is useful where the study has a fixed and limited number of cases, and therefore maximizing efficiency is critical. This may arise from either a particularly rare exposure or a particularly rare outcome. For example, in the 1960s some eight cases of vaginal adenocarcinoma were diagnosed in young women in Boston, Massachusetts. Vaginal cancer was previously virtually unknown in young women, and this particular disease was of an unusual histological type. To study the possible causes, the most efficient design is one in which causal factors are assessed using comparison subjects who are closely matched for the main confounders. To study the aetiology of this condition, for each patient with vaginal adenocarcinoma, four comparison subjects were chosen who were matched on sex, date of birth (within 5 days), hospital of birth, and ward or private type of service [32]. The same consideration of maximizing efficiency also applies if collecting data from the study subjects is expensive, so that the number of subjects needs to be kept a small as possible.
p. 202 For example, if there is a set of subjects with stored blood or tissue ↳ samples, careful matching can be done so that fewer samples need to be used for expensive chemical or genetic tests. Thus to assess the relationships of breast cancer with plasma levels of carotenoids and other indicators, stored blood samples collected up to 10

years before diagnosis from over 70 000 subjects in three cohort studies in Sweden were used. For each of 624 breast cancer cases, two controls matched for age, date of blood sample, and sampling centre were selected, and so biochemical analyses were done only on these samples [33].

Disadvantages of matching

Matching is a technique that involves both practical and conceptual difficulties. It has several disadvantages compared with other methods of controlling confounding. In a conventional study with new subject recruitment, to obtain each appropriate matched comparison subject, several potential comparison subjects may have to be approached and initial information gathered, making the study more expensive and difficult to set up. If the subjects can be selected from an existing data bank with information already available on the matching factors, this is not a problem, and matching is more often used in those circumstances. The matched design is prone to loss of data; if one member of a matched pair does not participate in or complete the study, the whole matched group usually has to be excluded. A further disadvantage has been referred to, that the analysis becomes complex if other factors have to be considered or if the matching ratio is variable.

The most important disadvantage of matching is that the matching factor cannot itself be assessed in the analysis in terms of its relationship to the outcome. Therefore matching should be used only for factors that are known to be important risk factors and thus important confounders, and should not be used if it is necessary to assess the relationship between the matched factor and the outcome in the study. Therefore matching is inappropriate for an exploratory study to answer a general question as to what are the causes of the outcome in question.

The further disadvantages of matching are in its inappropriate use resulting in overmatching, which will now be discussed.

Unnecessary matching: ‘overmatching’

If the study involves matching on a factor which is not a true confounder, this is often referred to as ‘overmatching’, although ‘unnecessary matching’ is a clearer description. It may occur in two situations.
p. 203 Suppose in a case–control study the subjects are matched on a factor which is associated with the exposure, ↳ but which is not itself associated with the outcome; therefore it is not a confounding factor, and it is unnecessary to control its effects. Because the matching factor is associated with exposure, the controls are chosen in a way that will make them more similar to the cases in terms of exposure. The differences in exposure between potential case and control subjects in the source population will be reduced in the study subjects, and an (incorrect) unmatched analysis will lead to an underestimate of the true outcome–exposure association. If a correct analysis is done, stratifying for the matching variable, using multivariate methods, or using an analysis on matched sets, the result of the unnecessary matching will be to increase the proportion of all case–control sets that are concordant for exposure. As these sets do not contribute to the estimate of odds ratio, the study results will not be biased, but the study will be less efficient as fewer sets of observations are contributing to the results. In the extreme, if the matching factor is very closely associated with the exposure, the study will have no ability to assess the exposure. In a cohort study, unnecessary matching also leads to inefficiency, but the study results will still be valid as specific analytic methods are not needed.

As an example, consider a case–control study assessing the relationship between passive exposure to tobacco smoke and lung cancer. Should lung cancer subjects and controls be matched on their personal smoking experience? As this is a major risk factor and is associated with passive smoking, it needs to be controlled, and matching with a correct analysis would give a powerful study. Should subjects be matched for the size of their family, or for the number of fellow workers they have? These factors are not themselves risk factors for lung cancer, but may be associated with passive smoking; matching would be detrimental. If we wish to guard

against the possibility of family size being a risk factor, perhaps by being an indicator of other exposures, it would be better to deal with it by stratification or multivariate analysis. Then we have freedom to assess if it is a risk factor, or a confounding factor; if we match on family size, we do not have flexibility. The error here is in matching for a factor which is associated with the exposure, but is not an independent risk factor for the outcome and so is not a confounder. However, matching for this factor can make it a confounding factor, which then has to be controlled in the analysis.

Unnecessary matching will also occur if the matching factor is not a confounding factor because it is part of a causal pathway linking exposure and outcome. In this situation, different methods of analysis may not help, as the study is fundamentally flawed. Consider a case–control study to assess the causes of bladder cancer in a workforce where records of both a chemical exposure and previous bladder cytology are available for the
p. 204 employees; if the ↳ association between the chemical exposure and bladder cancer is to be assessed, should subjects be matched on prior cytology findings? If they are, and the true causal chain is

chemical exposure → abnormal cytology → bladder cancer,

a matched study will probably show no difference in chemical exposure. To conclude from such results that chemical exposure is not linked to bladder cancer is wrong. In such situations, the study design is not so much wrong, as misapplied. The design used in fact tests a different hypothesis: is chemical exposure related to bladder cancer irrespective of prior cytology findings? This may be a question worth asking, although probably only in specific subgroups; for example, if chemical exposure increases the risk of bladder cancer in subjects with previous abnormal cytology, it may indicate a tumour-promotion effect, increasing the risk of progression from abnormal cytology to invasive cancer. In these situations the interpretation cannot be made solely on the data; it requires assumptions regarding the causal model and is dependent on whether the third factor is a confounding factor or not. If subjects are chosen without matching, and information on the third factor is collected, analyses can be done with and without control for that factor, and its associations with exposure and outcome can be assessed. If matching has been used, this flexibility is lost. The risks of unnecessary matching again show that matching should be used only after careful consideration, including knowledge about the confounding factors relevant to a given situation.

Another situation of overmatching is where matching is done on a factor which is itself affected by either the exposure or outcome in the study. This includes symptoms and signs produced either by the exposure or the outcome, including early indications of disease. Again, the issue is that these factors are not true confounders. For example, in a cohort study of smokers, if smokers and non-smokers are matched in terms of a history of cough, the effect of smoking on the subsequent incidence of lung cancer would be underestimated. This situation particularly applies to studies of the effects of drugs, where the dominant confounder is the underlying disease for which the drugs are prescribed. This is referred to as *confounding by indication*. A strong association between a drug and a disease may arise if the drugs are prescribed for clinical signs or symptoms which are associated with the early stages of the disease, so-called *protopathic bias*. For example, in the assessment of the association between aspirin use and Reye's syndrome (an encephalitis-like illness mainly affecting children, usually after a viral illness), such an association could arise because aspirin was prescribed
p. 205 to children who had symptoms ↳ which were early manifestations of Reye's syndrome. This is reverse causation: the putative outcome (disease) causes the exposure. A further case–control study concluded that the association was not explicable by such a mechanism, and strengthened the case for aspirin use being causal [34].

In summary, there are two major types of matching. Frequency matching is used to produce approximately similar distributions of key confounding factors in the groups being compared. It should be considered simply as a method of improving study efficiency, and the matching factors and other confounders should be dealt with in the analysis by the usual stratification and multivariate methods. Specific individual matching is of particular value for complex confounders which cannot otherwise be dealt with, and situations where

maximum efficiency is a priority. Such matching should only be carried out for factors which are true confounders. There should be good evidence that the matching factor is a strong risk factor for the outcome, with in addition an association with the exposure. Overmatching is produced if matching is carried out on a factor which is not a true confounder. Matching for a factor associated only with exposure will decrease efficiency and, if not appropriately analysed, will damage validity. It is important to avoid matching on a factor which is a component in the causal chain linking exposure and outcome, or on a factor which is affected by the exposure or the outcome, such as symptoms or signs.

Multivariate methods

The final method of controlling confounding is to analyse the data using a mathematical model that has the outcome as the dependent variable, and includes both the postulated causal factor and confounding factors in the equation. Multivariate analysis is a major subject in itself, and the purpose of this section is to present the key principles which are important in interpreting the results of these analyses.

Linear regression

While linear regression is not widely used in these studies, it is the simplest model from which the others are derived. For continuous variables, linear multiple regression is often appropriate, and is a standard technique included in most statistical texts and computer programs. For example, consider the assessment of whether maternal pre-pregnancy weight is related to the birthweight of the baby, taking account of any relationship with the mother's height. A linear multiple regression model could be used, where the dependent variable is
p. 206 birthweight, and the model includes maternal pre-pregnancy ↳ weight as one independent (or 'predictor') variable and height as another. Standard multiple regression methods produce a coefficient for the maternal weight variable which shows the relationship between it and birthweight, independent of height.

The mathematical expression is:

$$y = a + b_1 x_1 + b_2 x_2$$

where y is the outcome, birthweight, and is the dependent variable in the equation, x_1 is the mother's pre-pregnancy weight, x_2 is the mother's height, a is a constant with no intuitive meaning (it is the birthweight if x_1 and x_2 are both zero), and b_1 and b_2 are the regression coefficients. These coefficients are calculated to be the values that give the best fit of the equation set out above to the observed data. (Terminology varies: y can also be called the outcome variable or regressand the x variables can be called predictor variables, covariates, or regressors).

This simple mathematical equation makes several assumptions. For example, it assumes that the change in birthweight with the change in the mother's pre-pregnancy weight is linear, and therefore the numerical value b_1 represents the amount of change in birthweight associated with a change in the mother's pre-pregnancy weight of one unit. A similar linear assumption holds for b_2. The equation also assumes that the change in birthweight with pre-pregnancy weight is the same irrespective of the value of the mothers' height, i.e. there is no *interaction* between these two variables. These are assumptions inherent in the mathematical form of the equation. If the variable y represents 'risk', the coefficient b_1 with this linear model show the *difference* in risk associated with x_1, and independence of the effects of several variables means independence on a linear scale, i.e. the risk differences associated with each variable are additive. There are usually other assumptions involved

in the ways in which the coefficients are calculated; for example, the usual method of calculation will make the assumption that the variable *y* has a normal distribution.

Log-linear models: Poisson regression

In many health situations the main variables are not continuous, or even if they are continuous, such as an incidence rate, they have limits; a rate cannot be negative. Outcomes are often dichotomous (binary): diseased or not diseased, cured or not cured. Exposures may be continuous, dichotomous, or have several categories (e.g. age, sex, stage of disease), and so standard linear regression will not be appropriate. Many other models have been developed and applied to these situations.

p. 207 One model is to take the natural logarithm of the outcome variable, shown by ln *y*, and use the same equation as for linear regression but with ln *y*:

$$\ln y = a + b_1 x_1 + b_2 x_2$$

This transformation means that *y* cannot be negative, as negative numbers do not have logarithms (ln 0 = 1). This is a log-linear model; the right-hand side is linear, and the dependent variable has a logarithmic transformation. It is equivalent to an exponential model:

$$y = \exp(a + b_1 x_1 + b_2 x_2)$$

This model is widely used in the analysis of epidemiological cohort studies. The distribution of the dependent variable *y* determines the best method of fitting the model to the data. In epidemiological cohort studies, where the frequency of positive outcomes is low, a model using a Poisson distribution is often suitable; this method is called *Poisson regression*. The interpretation of the results of any log-linear model is the same.

As these models use a logarithmic, or exponential, model, the interpretation of the coefficients is different from a linear model. Consider a simple model with only one independent variable x_1 on the right-hand side. The coefficient b_1, for variable x_1, estimates the change in ln *y* for a change of one unit in x_1. If $x_1 = 1$ means that the subject is exposed, and $x_1 = 0$ means that they are unexposed, the change in outcome associated with x_1 is given by the following. Let y_e be the risk in the exposed group and y_u be the risk in the unexposed group. If $x_1 = 0$ (unexposed), $\ln y_u = a$ (as $x_1 = 0$, $b_1 x_1 = 0$). If $x_1 = 1$ (exposed), $\ln y_e = a + b_1$. Therefore the difference in the equations is b_1, which must equal $\ln y_e - \ln y_u$, and this equals $\ln(y_e/y_u)$. Hence

$$b_1 = \ln(y_e/y_u)$$

and

$$\exp(b_1) = (y_e/y_u) = \text{risk in exposed/risk in unexposed}$$

Therefore in a log-linear model with a binary exposure variable, the exponential of a coefficient equals the *risk ratio (relative risk)* associated with that exposure. For a continuous exposure variable, the exponential of a coefficient represents the risk ratio associated with a one unit change in the exposure factor. This contrasts with the linear model described earlier, where the coefficient showed the risk difference associated with the factor. It follows that where we have two or more independent variables in a log-linear model, the definition of
p. 208 independent effects is that the risk ratio for each factor is ↳ constant, and so the joint effect of two 'independent' factors is additivity with regard to the log of disease risk, which is *multiplicity* in the effect on the absolute disease risk.

Logistic regression

A further slightly different model is the most appropriate for case–control studies, although it can be used in any study design. It uses a further transformation of *y*. In the *logistic model* the logarithm of the odds, called the *logit* of disease risk, is used as the dependent variable in a linear regression equation. If *P* is the proportion of subjects in the study who have the outcome, or equivalently the probability that a randomly selected subject has the outcome, the logit of *P* is defined as ln[*P*/(1 − *P*)] where ln means the natural logarithm (to base *e*).

The logistic regression equation has the form

$$\ln [P/(1 - P)] = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots$$

where as before the *x* variables represent exposure factors and confounders, and the *b* terms are their coefficients. If an *x* variable is a numerical value (e.g. height), the *b* coefficient gives the change in logit *P* associated with a change of one unit of *x*, with the assumption of a linear relationship between the two. If *x* is binary, with the values of 0 or 1, the equation yields the odds ratio associated with *x*. For example, for sex we might define a variable so that $x_1 = 1$ for males and $x_1 = 0$ for females. For females, $x_1 = 0$, and so $b_1 x_1$ is zero. For males, $x_1 = 1$, and so the equation for males differs from that for females by having the extra term $b_1 x_1$, which is equal to b_1. The rest of the equation is the same. Therefore the difference in the logit *P* between males and females will be b_1:

$$\ln\left(\frac{P_m}{1-P_m}\right) - \ln\left(\frac{P_f}{1-P_f}\right) = b_1$$

where P_m is the risk for males and P_f is the risk for females. For two numbers *r* and *s*,

$$\ln r - \ln s = \ln (r/s)$$

Therefore

$$\ln\left[\frac{P_m/(1-P_m)}{P_f/(1-P_f)}\right] = b_1$$

p. 209 where $P_m/(1 - P_m)$ is the odds of the outcome in males and $P_f/(1 - P_f)$ is the odds in females. Therefore the quantity in the brackets is the *odds ratio* comparing male and female subjects:

$$\ln(\text{OR}) = b_1$$

and

$$\text{OR} = \exp(b_1)$$

Thus with the logistic regression model, the exponential of a coefficient b equals the odds ratio associated with the variable x, if this is a binary variable coded as 0 or 1. If x is continuous, b gives the odds ratio associated with a change in x of one unit.

A simple example of this is given by the comparison between tolbutamide and placebo, using the data shown in Ex. 6.18. These data could be analysed by any log-linear model. We will apply a logistic model, as was done in the original analysis:

$$\ln [P/(1 - P)] = a + b_1 x_1$$

where P is the probability of death, and x_1 is 1 for tolbutamide and zero for placebo. For the placebo group

$$P_p = 21/205 = 0.1024$$

(from Ex. 6.18, Subtable A). Therefore

$$\text{logit } P_p = \ln (0.1024/0.8976) = -2.171 = a.$$

For the tolbutamide group

$$P_t = 30/204 = 0.1471.$$

Therefore

$$\text{logit } P_1 = \ln(0.1471/0.8529) = -1.758 = a + b_1.$$

Thus

$$b_1 = -1.758 - (-2.171) = 0.413$$

and

$$\exp(b_1) = \exp(0.413) = 1.511.$$

p. 210 Therefore the odds ratio is 1.511. We can also calculate the odds ratio from Ex. 6.18, Subtable A, by the usual method, with the same result: odds ratio = (30 × 184)/(21 × 174) = 1.511.

To account for other factors, the model becomes

$$\ln[P/(1-P)] = a + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots$$

where the other x variables represent other factors. The value of multivariate analysis is that many other factors can be included in the equation. In the analysis of this study, a logistic model was used with 15 other factors (the nine factors shown in Ex. 6.12, plus sex, race, systolic blood pressure, diastolic blood pressure, visual acuity, and creatinine level). With all these other factors included, the value of b_1 in the presence of these other factors was 0.40, giving an odds ratio of 1.49 for tolbutamide. Thus although there were some considerable differences between the tolbutamide and placebo groups, as shown in Ex. 6.12, these did not in aggregate produce any great difference in the main result; the odds ratio for tolbutamide versus placebo only changed from 1.51 to 1.49. The results can also be expressed as the difference in mortality rates, as was done in the original paper [10].

Use of dummy variables

Frequently, factors with several categories are relevant, and these are often best handled by using a number of 'dummy' binary variables to represent all the categories. **Exhibit** 6.27 gives some results using this method. The data are from a case–control study comparing 83 patients with malignant melanoma with 83 controls chosen from the general population [35], and show the results for just two factors, the number of palpable moles (naevi) on the upper arm assessed by an interviewer (three categories) and the response to a question on whether the subject had ever had a severe sunburn (two categories). The results from cross-tabulations showed strong associations with both the number of moles and a history of sunburn. However, several other factors were associated with these two features and are also related to melanoma, which therefore are confounding factors. These included the severity of skin freckling (three categories), the usual reaction to sun exposure (four categories), and hair colour (three categories). Control for each of these confounders singly can be done fairly easily by cross-tabulations. However, to assess the relationship between moles and melanoma, with control for freckles, sun reaction, hair colour, and sunburn simultaneously by cross-tabulations, would mean that 3 × 4 × 3 × 2 = 72 separate tables showing the case–control distribution by numbers of moles would ↳ have to be generated, and in this small study many of these tables would have few or no observations.

p. 211

MULTIVARIATE ANALYSIS

A: Cross tabulations

	No. of moles on upper arm			History of sunburn	
	0	1–2	3 +	No	Yes
Cases, number	32	16	35	34	49
Controls, number	62	17	4	57	26
Odds ratio	1.0 (R)	1.82	16.95	1.0 (R)	3.16
B: Logistic regression with one factor only					
Coefficient *b*		0.6008	2.830		1.150
Exp(*b*) = odds ratio		1.82	16.95		3.16
C: Logistic regression with both factors, plus quantity of freckles (3 categories), reaction to sun exposure (4 categories), and hair colour (3 categories)					
Coefficient *b*		0.3011	2.587		0.4276
Exp(*b*) = odds ratio		1.35	13.29		1.53

Ex. 6.27. Multivariate analysis: results from a case–control study comparing 83 patients with malignant melanoma with 83 controls from the general population. Results are shown for two factors, number of moles on the upper arm and history of severe sunburn, derived by (A) cross-tabulation, (B) a logistic regression fitting only the one factor, and (C) a logistic regression fitting five factors, represented by 10 binary variables. From Elwood *et al*. [35]; fuller results are given in that paper

Therefore it is more useful to include these five factors in a logistic regression equation, expressing each factor as a number of dummy variables, the number being 1 less than the number of categories. Thus for the number of moles, one binary variable was used with the value 1 for subjects who had one to two moles on the upper arm, and zero otherwise, and another with the value 1 was used for subjects with three or more moles, and zero otherwise. Where both these factors are zero the equation gives the risk in the referent category—subjects with no moles. If a logistic regression is fitted with just one factor, the results will be identical to a simple cross-tabulation, and the exponential of the coefficient *b* will be equal to the odds ratio obtained by the usual calculation on the simple table. These results are shown in Ex. 6.27, part B. Then if a model is fitted which includes the variables representing all the factors listed above, the coefficient *b* for each variable will give the odds ratio associated with that variable, controlled for the effects of all the other variables in the equation. Exhibit 6.27, part C, shows that with these coefficients, ↳ obtained from a model with 10 variables, the odds ratios for number of moles are still high, whereas the odds ratio for history of sunburn is 1.53, considerably lower than the crude odds ratio of 3.16. This shows that most of the association with sunburn seen in the

p. 212

simple analysis is not causal, but is produced by confounding by one or more of the other factors included in the equation. To determine which factor, analyses can be done fitting each confounding factor singly and seeing how the odds ratio for sunburn changes. The program used to calculate the coefficients will also allow the estimation of the statistical significance of these adjusted coefficients, and this aspect of multivariate analysis is considered further with the same example in Chapter 7 (Ex. 7.12), p. 250.

The logistic model is the multivariate model most widely used in epidemiological studies and clinical trials, but many other models exist. To use such models the confounding factors must be recognized in advance, and quantitative information on them must be collected. Clearly, considerable care is necessary to use such models properly.

Limitations of multivariate analysis

Multivariate analysis can deal with only a limited number of factors. Its scope is constrained by the number of study subjects, which should be much greater than the number of variables included. Most computer programs for such analyses have a limit to the number of factors they can deal with satisfactorily. Factors can be used in different ways. For example, a factor like age can be entered as one continuous variable or represented by a number of categories, each with a corresponding variable. If interactions are to be examined, extra interaction terms must be used. Thus for studies which include data on many factors, the most relevant must be selected before a multivariate model is applied. The next section on the use of the principles of confounding in analysis will be helpful in this regard.

Multivariate analysis of matched data: conditional models

Multivariate analysis for individually matched studies uses methods that take account of the matching, i.e. they consider each case and its matched control(s) as a set. These methods are referred to as 'conditional' models. They require skilled application. The results are presented in the same way as has been shown, with the same interpretation of the coefficients. Further information is given in texts [36,37].

Factors to be included in a multivariate model should be studied in detail, and issues such as the distribution of
p. 213 the observations, the need for ↳ transformations, and the appropriateness of assumptions such as a linear relationship to risk should be considered. Multivariate analysis is best regarded as a powerful but complex and demanding type of analysis, appropriate to the final stages of analysis of a study, rather than as a magical black box to provide a short cut to a result.

3. Other applications of confounding: Use of the definition of confounding in designing a study

Now that we have discussed confounding and the methods available to control it, it is useful to go back to the definition of confounding and see how this can assist us in the *design* of studies. The definition of confounding provides an answer to the problem of feeling that because there are so many possible confounding factors, no satisfactory study can be designed.

In designing any study, we should make a list of the factors that are likely to be associated with the *exposure* under study, and another list of factors that are likely to be associated with the *outcome*. To do this, we may need to review the literature and consult reference works and people with specialized knowledge. Any factor that appears on both of these lists is a potential confounding factor, and we need to plan how to deal with it. Factors that appear on only one list, but which are likely to have a very strong association with either outcome or exposure, may be prudently included as potential confounding factors, as even a small difference in their

distribution between the groups being compared may be sufficient to introduce confounding. The use of this approach will often reduce an apparently infinite number of potential confounders to a finite, and often fairly small, list of specific factors.

The options available for confounder control (Ex. 6.10) can then be considered. In practice, many potential confounders will not in fact be confounding, in that their associations with outcome and exposure in the study data are often weak and unimportant, but this will be known only if data on the confounders are collected. The most commonly used approach in non-randomized studies is to apply some restriction and to collect data on the potential confounders to allow stratification or multivariate methods to be used in the analysis. Individual matching should be used only when there are specific advantages to it.

In randomized studies only one list, of factors likely to be related to the outcome, is needed, and data on these should be collected, where possible, to assess whether the groups are in fact comparable on these factors. In large-scale ↳ randomized studies, for example of population interventions, samples of the groups may be selected for this purpose.

p. 214

Use of the definition of confounding in planning an analysis

Similarly, there are often a large number of potential confounding factors in the analysis. Initial data analysis can be used to decide if any of these factors are related to both the outcome and the exposure. This can be done by generating two sets of cross-tabulations: between each of the factors and outcome, and between each of the factors and exposure. The association between the potential confounding factor and the outcome should be examined within the non-exposed group, and the association between the potential confounding factor and exposure should be examined within the group without the outcome under consideration. Only those factors which show associations with both exposure and outcome need to be considered further as confounding factors. It is important to emphasize that it is the strength of the association, measured by the odds ratio or other measure, that matters, not its statistical significance. This type of initial analysis will often reduce a formidably large data set to a much simpler situation with only a few major confounding factors, and these can be analysed further by stratification or multivariate methods.

An alternative method of deciding which factors are confounding is often easier, and relies on the fact that confounding is demonstrated if a stratified or multivariate analysis is performed and the unconfounded relative risk or odds ratio estimate is different from the crude estimate. Therefore a practical approach to the analysis of a large data set is first to produce the basic table comparing exposure with outcome, and calculate the crude odds ratio or relative risk. Then each potential confounder is considered in a reasonable number of categories (five are usually sufficient), and stratified or multivariate analyses performed including the potential confounders, calculating the adjusted odds ratio or relative risk. If this adjusted ratio is similar to the crude ratio, there is no substantial confounding by that variable. In stratified analysis, the Mantel–Haenszel odds ratio or relative risk is usually used as the adjusted (unconfounded) estimate. Using multivariate methods, if a factor or a set of factors is confounding, the coefficient and therefore the odds ratio estimate for the exposure–outcome association will change when the confounding factors are added into the equation. Multivariate methods are very useful, as sets of potential confounding factors can easily be used and factors shown to have a confounding effect can be kept in the model while other ↳ factors are tested. The use of multivariate models to deal with confounding requires a close attention to the meaning of the various factors with regard to the causal hypothesis that is being assessed.

p. 215

It is the size of the difference between the crude and adjusted risk ratio, not its statistical significance, which indicates confounding. This approach is different from the methods of selecting factors in multivariate models used in standard computer routines, such as forward and backward selection methods, as these are based on the statistical significance of each variable. In a large study with many variables, a strategy of setting a

maximum acceptable change in the odds ratio or relative risk, such as 10 or 5 per cent, can be applied to potential confounders individually or in groups [38]. In published studies, a statement is sometimes given that many specified potential confounders were assessed, but that none of them changed the risk ratio estimate by more than a given percentage, and so they were not included in the final analysis. Models can be fitted including potential confounders cumulatively, or including all potential confounders and then excluding some; the critical change is the odds ratio for the main association being studied. Thus a large number of potential confounders can be reduced to a small number of actual confounders. Two or more potential confounders considered together may have a confounding effect, even if singly they do not; this is easily assessed using multivariate models. Therefore a reasonable approach is to proceed further with a reduced data set, keeping factors that have shown a confounding effect on simpler analysis plus those of major predetermined importance. This data set should then present a less forbidding challenge.

Mendelian randomization: a special situation allowing control of confounding

The term ‘Mendelian randomization’ has been used for studies that exploit the random assortment of alleles at the time of gamete formation to give a method of avoiding confounding. The critical issue is that the random assortment of alleles should be independent of the distribution of behavioural and environmental factors in the population. Thus if a specific genetic allele is associated with a particular causal mechanism, the mechanism can be assessed on the basis of this allele, which is likely to be distributed independently of other factors. Thus an analysis based on Mendelian randomization may have advantages similar to a randomized clinical trial [39,40]. However, there are several limitations of this approach [41], and only a few examples as yet of its application.

p. 216 What sparked interest in this issue was that an association was seen between low cholesterol levels and increased cancer rates in observational studies and in some trials of cholesterol-lowering agents. It was suggested that this could be either a causal effect, with a reduction in cholesterol causing an increase in cancer risk (which would require the treatment of high cholesterol to be reconsidered), or be due to presymptomatic cancers causing a reduction in cholesterol levels (protopathic bias). Conventional observational studies of this association will be limited by the many possible factors associated with cholesterol levels (such as other dietary factors). Katan [42,43] proposed comparing cancer risks in people with different polymorphisms of the apolipoprotein E gene. Individuals with the E2 allele have lower levels of cholesterol because their genotype gives them greater efficiency in removing cholesterol from plasma. Therefore if low cholesterol causes an increased risk of cancer, people with the E2 allele should have higher cancer risks, and a comparison of subjects with different gene types should be free of confounding as the gene type is distributed randomly.

In an example, many cohort and case–control studies have shown that people with a high intake of green cruciferous vegetables (such as broccoli and cabbage) show lower risks of lung cancer. This may be because these vegetables contain isothiocyanates, which may have a protective effect against cancer. Two genes, *GSTM1* and *GSTT1*, are implicated in the production of the enzyme glutathione-S-transferase, which is thought to eliminate isothiocyanates. Subjects who have ‘negative’ alleles of these two genes do not produce glutathione-S-transferase, and so have a lesser ability to metabolize isothiocyanates. If this mechanism applies, these subjects should show a greater protective effect from a high intake of cruciferous vegetables than subjects with ‘positive’ alleles who do produce the enzyme. In a large case–control study of lung cancer in six East European countries, a dietary questionnaire was used and a blood sample was taken on which genotyping for the two genes was done. In subjects who were negative for one or both of the key alleles, a high consumption of cruciferous vegetables protected against lung cancer; the odds ratio for those who had both gene alleles negative was 0.28. No major protective effect was seen in people who were positive for the two genes: odds ratio 0.88 [44]. This gives support to the proposed mechanism. The advantage of using the logic of Mendelian

randomization is that the presence of the relevant alleles should be independent of the many factors that could be confounding in a conventional dietary study. In this study it was shown that gene type was not associated with factors such as age, country, smoking, education, or dietary factors including cruciferous vegetable consumption.

The limits of confounder control

p. 217

A central issue in the interpretation of observational studies is whether the methods of control for confounding, which we have described in this chapter, can ever be good enough to give complete assurance that confounding has been controlled. On the one hand, studies such as the case–control studies and subsequently the prospective cohort studies of cigarette smoking and lung cancer have provided strong and consistent evidence demonstrating a causal relationship, and explaining it in quantitative terms, so that the strength of the available knowledge on that topic can be regarded as equivalent to the strength of evidence produced by randomized controlled trials in other contexts. Indeed, the aim of those carrying out epidemiological studies using observational methods is often stated in these terms—to be able to come to a similar level of confidence in the results as can be achieved by the randomized trial method. The counter-argument is that no observational study can control for a confounder that is not specifically included in the study, and even for those factors included, misclassification errors limit the ability to control their confounding effects.

The key advantage of the randomized trial method is its ability to reasonably exclude confounding, not only by factors that have been measured and assessed, but also by other factors. The real weakness of observational studies is that they provide no protection against the confounding effects of other factors that have not been measured, and may not be recognized or known. The methods described in this chapter can deal with confounding by factors that can be identified and measured, although even for these, confounding may not be completely controlled because of observational errors, limitations in how confounding factors are defined and measured, and statistical issues.

Several recent examples have been particularly influential in emphasizing the limitations of observational studies, even when well performed and analysed. These are situations where after extensive observational studies produced consistent results, large-scale randomized trials have produced quite different results. An example is the study of the potential protective effect of beta-carotene, contained in many vegetables. Extensive observational studies, both case–control studies and long-term prospective cohort studies, using the best available methods of dietary assessment and carried out by some of the world's leading research groups, produced generally consistent evidence that subjects with higher serum retinol or beta-carotene levels or with higher beta-carotene intake had substantially reduced cancer risks. This was supported by laboratory evidence that beta-carotene and its metabolites had an anti-carcinogenic action in animals. This evidence was collated over many years and a consensus formed that it was strong and consistent enough to justify intervention
p. 218 studies ↳ using beta-carotene dietary supplements to actively prevent many cancers [45]. Although the evidence for benefit was substantial, as Peto *et al*. [45] stated: 'Preventive measures, especially those which may be relevant over a long period to many million people, deserve particularly rigorous evaluation'. Large clinical trials were set up, but in contrast with all the previous evidence, these trials either showed no beneficial effect, or showed higher cancer rates in those receiving the beta-carotene supplements, and such trials had to be stopped promptly [46,47]. 'The cancer prevention community was stunned' by these results [48].

Another example relates to a reduction in the risk of coronary heart disease in women using oestrogen and/or progesterone hormones. Several observational studies showed that hormone users had a substantially reduced risk of coronary heart disease, with either no change in the risk of stroke or a slight reduction [49]. Therefore hormone therapies became widely recommended as a protection against heart disease in women. A trial

involving over 16 000 post-menopausal women at 40 centres in the USA, the Women's Health Initiative (WHI) trial, was set up to validate this, but it showed an increased rate of heart disease and stroke in women randomized to hormone therapy (combined oestrogen–progesterone) [50]. As a result, this arm of the WHI trial was terminated early. Similar results were shown in other randomized trials. These clinical trials changed practice and destroyed the case for the benefits of hormone therapy that had been built up through observational studies. It has been pointed out that the confounding effects of socio-economic status had been inadequately dealt with in the observational studies and also that an analysis of observational studies taking account of the changes over time in the exposure (hormone use) changed the results, and these two influences together could explain the discrepancies between the observational studies and the clinical trials [51]; this commentary concluded by stating 'however, observational studies are not a substitute for clinical trials no matter how sophisticated the statistical adjustments may seem'.

Another example is that antioxidant vitamins such as vitamin C have shown protective effects against cardiovascular disease, cancer, and total mortality in many observational studies, but well-conducted randomized controlled trials of supplements with antioxidants do not show any beneficial effects. The largest observational study shows protective effects of high vitamin C plasma levels on coronary heart disease mortality, with odds ratios of 0.70 in men and 0.63 in women. This study controlled for age, systolic blood pressure, cholesterol, body mass index, smoking, diabetes, and vitamin supplement use. However, the largest randomized trial set up to verify this association showed a small increase in risk associated with vitamin C
p. 219 supplementation, with a relative ↳ risk of 1.06 [52]. This situation must be due to the influence of unmeasured confounding factors in the observational studies.

The challenge is the conflict between the desire to have the best quality scientific evidence and avoid premature action on the basis of apparently strong observational studies, and the reality that requiring a randomized trial may delay action for many years, or in some circumstances a randomized trial may never be done. For example, the definitive randomized trial of the preventive action of folic acid on birth defects is described in detail in Chapter 11. In this situation, an observational study completed 10 years before the randomized trial was completed in fact gave the correct answer, and public health action at that time would have prevented many birth defects which otherwise occurred. Another example is the current situation with screening by routine skin examinations with the aim of reducing deaths from melanoma, a dangerous skin cancer. Although authorities generally agree that a screening programme should only be instituted on the basis of randomized trial evidence, no randomized trial of skin screening has been done anywhere in the world, despite a successful pilot study [53], and none may ever be done because of the cost, the numbers of subjects, and the time required.

Self-test questions (answers on p. 499)

Q6.1 Suppose an as yet unidentified dietary constituent (X) greatly reduces the risk of cancer. It is distributed in foodstuffs in a similar way to an easily measured dietary constituent (D) which itself has no effect on cancer incidence. In a prospective cohort study, dietary intakes of D are measured, and subsequent cancer incidence recorded. What will be the result?

Q6.2 In the same situation as in Q6.1, an intervention trial using pure compound D is carried out. What will the result be?

Q6.3 An innovative pre-school reading programme is launched by asking teachers to volunteer to trial the programme, and the children in the programme are then compared with other children in the same school systems. What are the results likely to show?

Q6.4 Exhibit 6.12 compares the characteristics of two groups in a randomized trial. Why are tests of the significance of the differences between the two groups not presented?

Q6.5 The experience of treating a disease, which can present either early or late, is compared between two hospitals. In hospital A, the success rates are 40 per cent for early disease and 25 per cent for late disease, based p. 220 ↳ on 500 and 100 patients in each category; in hospital B, the success rates are 60 per cent for early disease and 40 per cent for late disease, based on 100 and 200 patients, respectively. How does the performance of hospital B compare with hospital A for each stage of disease; and how does this compare with the crude comparison based on all patients treated in each hospital?

Q6.6 Calculate the Mantel–Haenszel measure of relative risk for the data given in Q6.5.

Q6.7 Again using the example given in Q6.5, assume that over the whole country 25 per cent of patients with this condition are early at presentation. Using a standard population of 25 per cent early and 75 per cent late disease, calculate the direct standardized success rates for each of hospitals A and B. What is the ratio of these direct standardized rates?

Q6.8 In a case–control study, the role of previous injury in producing arthritis of the knee is assessed. In men, 300 of 900 men with arthritis of the knee reported previous injury, compared with 100 of 400 male controls. For women, four of 44 female cases reported previous injury, compared with 50 of 450 female controls. Calculate the overall crude odds ratio, the sex-specific odds ratios, and the Mantel–Haenszel odds ratio.

Q6.9 In an individually pair-matched case–control study, 200 pairs are concordant for exposure and 100 pairs are concordant for lack of exposure; for 200 pairs, the case is exposed and the control is not; for 50 pairs, the control is exposed and the case is not. What is the odds ratio? What would the odds ratio be if an analysis ignoring the matching was performed?

Q6.10 In a multivariate analysis, a binary variable represents oral contraceptive use, being coded 1 for ever use and zero for never use, in a prospective study assessing cardiovascular disease. On fitting ever use of oral contraceptives, plus age, the coefficient is 0.45. When the subject's weight is added to the equation, this coefficient changes to become −0.08. What is the odds ratio for the association with ever use of oral contraceptives, and what is the confounding effect of weight?

References

1. Charig CR, Webb DR, Payne SR, Wickham JEA. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *BMJ* 1986; **292**: 879–882. 10.1136/bmj.292.6524.879
WorldCat Crossref

p. 221 2. Charig CR. Confounding and Simpson's paradox: multiple regression would confound the clinicians. *BMJ* 1995; **310**: 329. 10.1136/bmj.310.6975.329b
WorldCat Crossref

3. Simpson EH. The interpretation of interaction in contingency tables. *J R Statist Soc* 1951; **2**: 238–241.
WorldCat

4. Paffenbarger RS, Hale WE. Work activity and coronary heart mortality. *N Engl J Med* 1975; **292**: 545–550. 10.1056/NEJM197503132921101
WorldCat Crossref

5. Adish AA, Esrey SA, Gyorkos TW, Jean-Baptiste J, Rojhani A. Effect of consumption of food cooked in iron pots on iron status and growth of young children: a randomised trial. *Lancet* 1999; **353**: 712–716. 10.1016/S0140-6736(98)04450-X
WorldCat Crossref

6. Altman DG, Schulz KF, Moher D, *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134**: 663–694.
WorldCat

7. Dodd JM, Hedayati H, Pearce E, Hotham N, Crowther CA. Rectal analgesia for the relief of perineal pain after childbirth: a randomised controlled trial of diclofenac suppositories. *BJOG* 2004; **111**: 1059–1064. 10.1111/j.1471-0528.2004.00156.x
WorldCat Crossref

8. Cook CC, Scannell TD, Lipsedge MS. Another trial that failed. *Lancet* 1988; **i**: 524–525. 10.1016/S0140-6736(88)91309-8
WorldCat Crossref

9. English DR, Burton RC, Del Mar CB, Donovan RJ, Ireland PD, Emery G. Evaluation of aid to diagnosis of pigmented skin lesions in general practice: controlled trial randomised by practice. *BMJ* 2003; **327**: 375–380. 10.1136/bmj.327.7411.375
WorldCat Crossref

10. University Group *Diabetes* Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II: mortality results. *Diabetes* 1970; **19**(Suppl.2): 785–830.
WorldCat

11. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 1980; **303**: 1038–1041. 10.1056/NEJM198010303031804
WorldCat Crossref

12. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen- year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982; **69**: 349–355.
WorldCat

13. Bruner AB, Joffe A, Duggan AK, Casella JF, Brandt J. Randomised study of cognitive effects of iron supplementation in non-anaemic iron-deficient adolescent girls. *Lancet* 1996; **348**: 992–996. 10.1016/S0140-6736(96)02341-0
WorldCat Crossref

14. Ashby D. Can iron supplementation improve cognitive functioning? *Lancet* 1996; **348**: 973. 10.1016/S0140-6736(05)64919-7

WorldCat Crossref

15. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003; **290**: 1624–1632. 10.1001/jama.290.12.1624
WorldCat Crossref

16. Labrie F, Candas B, DuPont A, *et al*. Screening decreases prostate cancer death: first analysis of the 1988 Quebec prospective randomized controlled trial. *Prostate* 1999; **38**: 83–91. 10.1002/(SICI)1097-0045(19990201)38:2⟨83::AID-PROS1⟩3.0.CO;2-B
WorldCat Crossref

17. Labrie F, Candas B, Cusan L, *et al*. Screening decreases prostate cancer mortality: 11-year follow-up of the 1988 Quebec prospective randomized controlled trial. *Prostate* 2004; **59**: 311–318. 10.1002/pros.20017
WorldCat Crossref

18. Boer R, Schroder FH. Quebec randomized controlled trial on prostate cancer screening shows no evidence for mortality reduction. *Prostate* 1999; **40**: 130–134. 10.1002/(SICI)1097-0045(19990701)40:2⟨130::AID-PROS9⟩3.0.CO;2-X
WorldCat Crossref

19. Elwood M. A misleading paper on prostate cancer screening. *Prostate* 2004; **61**: 372. 10.1002/pros.20160
WorldCat Crossref

20. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–748.
WorldCat

p. 222 21. International Agency for Research on Cancer. *Cancer Incidence in Five Continents*, Vol. VIII. Lyon, France: International Agency for Research on Cancer, 2002.
Google Scholar Google Preview WorldCat COPAC

22. Elwood JM, Gallagher RP, Davison J, Hill GB. Sunburn, suntan and the risk of cutaneous malignant melanoma: the Western Canada Melanoma Study. *Br J Cancer* 1985; **51**: 543–549. 10.1038/bjc.1985.77
WorldCat Crossref

23. Lyon JL, Gardner JW, West DW, Stanish WM, Hebertson RM. Smoking and carcinoma *in situ* of the uterine cervix. *Am J Public Health* 1983; **73**: 558–562. 10.2105/AJPH.73.5.558
Crossref

24. Karlberg J, Chong DS, Lai WY. Do men have a higher case fatality rate of severe acute respiratory syndrome than women do? *Am J Epidemiol* 2004; **159**: 229–231. 10.1093/aje/kwh056
WorldCat Crossref

25. Elwood JM. Wood exposure and smoking: association with cancer of the nasal cavity and paranasal sinuses in British Columbia. *Can Med Assoc J* 1981; **124**: 1573–1577.
WorldCat

26. Daly E, Vessey MP, Hawkins MM, Carson JL, Gough P, Marsh S. Risk of venous thromboembolism in users of hormone replacement therapy. *Lancet* 1996; **348**: 977–980. 10.1016/S0140-6736(96)07113-9
WorldCat Crossref

27. Halvorsen I, Andersen A, Heyerdahl S. Girls with anorexia nervosa as young adults. Self-reported and parent-reported emotional and behavioural problems compared with siblings. *Eur Child Adolesc Psychiatry* 2005; **14**: 397–406. 10.1007/s00787-005-0489-0
WorldCat Crossref

28. Scott WK, Zhang F, Stajich JM, Scott BL, Stacy MA, Vance JM. Family-based case–control study of cigarette smoking and Parkinson disease. *Neurology* 2005; **64**: 442–447. 10.1212/01.WNL.0000150905.93241.B2
WorldCat Crossref

29. Swerdlow AJ, De Stavola BL, Floderus B, *et al*. Risk factors for breast cancer at young ages in twins: an international population-based study. *J Natl Cancer Inst* 2002; **94**: 1238–1246. 10.1093/jnci/94.16.1238
WorldCat Crossref

30. Carlsson S, Hammar N, Grill V, Kaprio J. Alcohol consumption and the incidence of type 2 diabetes: a 20-year follow-up of the Finnish twin cohort study. *Diabetes Care* 2003; **26**: 2785–2790. 10.2337/diacare.26.10.2785
WorldCat Crossref

31. Cederlöf R, Jonsson E, Kaij L. Respiratory symptoms and 'angina pectoris' in twins with reference to smoking habits: an epidemiological study with mailed questionnaire. *Arch Environ Health* 1966; **13**: 726–737.
WorldCat

32. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med* 1971; **284**: 878–881. 10.1056/NEJM197104222841604
WorldCat Crossref

33. Hulten K, Van Kappel AL, Winkvist A, *et al*. Carotenoids, alpha-tocopherols, and retinol in plasma and breast cancer risk in northern Sweden. *Cancer Causes Control* 2001; **12**: 529–537. 10.1023/A:1011271222153
WorldCat Crossref

34. Forsyth BW, Horwitz RI, Acampora D, *et al*. New epidemiologic evidence confirming that bias does not explain the aspirin/Reye's syndrome association. *JAMA* 1989; **261**: 2517–2524. 10.1001/jama.261.17.2517
WorldCat Crossref

35. Elwood JM, Williamson C, Stapleton PJ. Malignant melanoma in relation to moles, pigmentation, and exposure to fluorescent and other lighting sources. *Br J Cancer* 1986; **53**: 65–74. 10.1038/bjc.1986.10
WorldCat Crossref

36. Rothman KJ, Greenland S. *Modern Epidemiology* (2nd edn). Philadelphia, PA: Lippincott–Raven, 1998.
Google Scholar Google Preview WorldCat COPAC

37. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford: Oxford Scientific Publications, 1993.
Google Scholar Google Preview WorldCat COPAC

38. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 1993; **138**: 923–936.
WorldCat

39. Hingorani A, Humphries S. Nature's randomised trials. *Lancet* 2005; **366**: 1906–1908. 10.1016/S0140-6736(05)67767-7
WorldCat Crossref

40. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003; **32**: 1–22. 10.1093/ije/dyg070
WorldCat Crossref

41. Brennan P. Commentary: Mendelian randomization and gene-environment interaction. *Int J Epidemiol* 2004; **33**: 17–21. 10.1093/ije/dyh033
WorldCat Crossref

42. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1986; **i**: 507–508. 10.1016/S0140-6736(86)92972-7
WorldCat Crossref

43. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Int J Epidemiol* 1986; **33**: 9 (reprinted from *Lancet* 1986; **i**: 507–508).
WorldCat

44. Brennan P, Hsu CC, Moullan N, *et al*. Effect of cruciferous vegetables on lung cancer in patients stratified by genetic status: a Mendelian randomisation approach. *Lancet* 2005; **366**: 1558–1560. 10.1016/S0140-6736(05)67628-3
WorldCat Crossref

45. Peto R, Doll R, Buckley JD, Sporn MB. Can dietary beta-carotene materially reduce human cancer rates? *Nature* 1981; **290**: 201–208. 10.1038/290201a0
WorldCat Crossref

46. The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 1994; **330**: 1029–1035. 10.1056/NEJM199404143301501
WorldCat Crossref

47. Omenn GS, Goodman GE, Thornquist MD, *et al*. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N Engl J Med* 1996; **334**: 1150–1155. 10.1056/NEJM199605023341802
WorldCat Crossref

48. Duffield-Lillico AJ, Begg CB. Reflections on the landmark studies of beta-carotene supplementation. *J Natl Cancer Inst* 2004; **96**: 1729–1731. 10.1093/jnci/djh344
WorldCat Crossref

49. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* 1991; **20**: 47–63. 10.1016/0091-7435(91)90006-P
WorldCat Crossref

50. Rossouw JE, Anderson GL, Prentice RL, *et al*. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002; **288**: 321–333. 10.1001/jama.288.3.321
WorldCat Crossref

51. Petitti DB, Freedman DA. Invited commentary: how far can epidemiologists get with statistical adjustment? *Am J Epidemiol* 2005; **162**: 415–418. 10.1093/aje/kwi224
WorldCat Crossref

52. Lawlor DA, Davey SG, Kundu D, Bruckdorfer KR, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet* 2004; **363**: 1724–1727. 10.1016/S0140-6736(04)16260-0
WorldCat Crossref

53. Aitken JF, Elwood JM, Lowe JB, Firman DW, Balanda KP, Ring IT. A randomised trial of population screening for melanoma. *J Med Screen* 2002; **9**: 33–37. ↳ 10.1136/jms.9.1.33
WorldCat Crossref