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

2 Study designs which can demonstrate and test causation

J. Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.003.02 Pages 19–52

Published: February 2007

Abstract

This chapter is divided into two parts. The first part describes the key study designs used in clinical and epidemiological studies of causation. The designs are defined by two features: whether the study subjects are selected by their exposure or by their outcome, and the time dimensions of the study. The second part describes the key strengths and weaknesses of each type of study, with many examples. Self test questions are provided at the end of the chapter.

Keywords: cause and effect, clinical studies, epidemiological studies, causation, time dimensions

Subject: Public Health, Epidemiology

Collection: Oxford Scholarship Online

> *All true and fruitful natural philosophy hath a double scale or ladder, ascendant and descendent, ascending from experiments to the inventions of causes, and descending from causes to the invention of new experiments.*

—Francis Bacon: The advancement of learning, II; 1605

In the first part of this chapter, we will describe the key study designs used in clinical and epidemiological studies of causation. The designs are defined by two features: whether the study subjects are selected by their exposure or by their outcome, and the time dimensions of the study. In the second part we will describe the key strengths and weaknesses of each type of study, with many examples.

Part 1. Types of study: The relationship of study design to the definition of causation

The definition of a causal factor given in Chapter 1 was ‘a factor whose operation increases the frequency of an event’. This implies that (i) people who are affected by the causal agent will have a higher frequency of the defined outcome, and (ii) individuals with the defined outcome will have a greater frequency of past exposure to the causal agent.

Thus there are two general types of comparative study. To test statement (i), we compare a group of people exposed to the putative causative factor with a group who are not exposed—a *cohort* study. The randomized trial introduced in Chapter 1 is a special type of cohort study, where the two groups are defined after randomization. To test statement (ii), we compare a group of people who have already experienced the outcome with a group of people without the outcome—a *case–control* study.

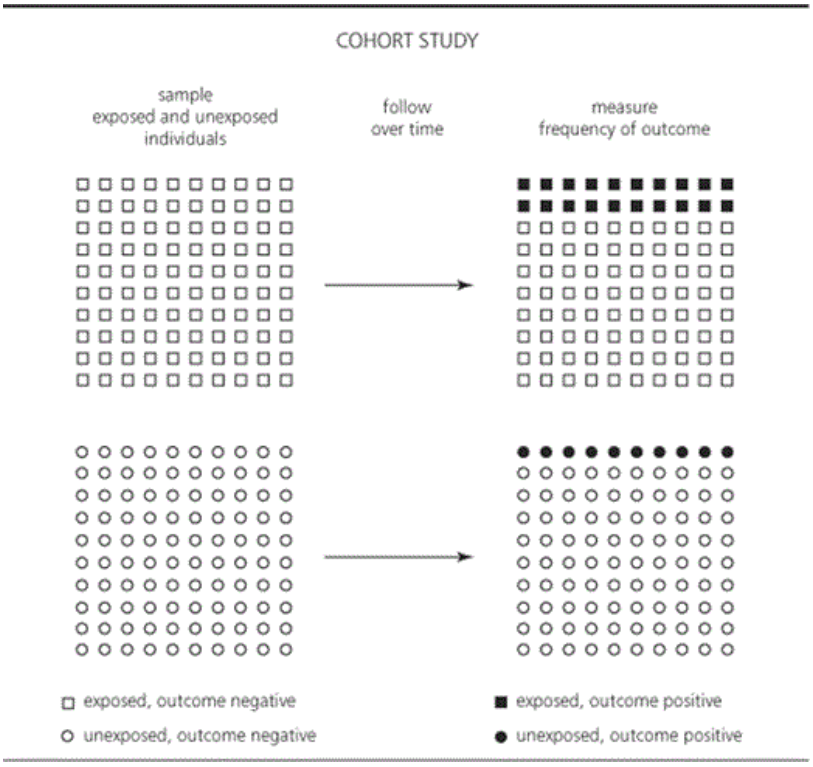
p. 20 These studies are set up to test causal hypotheses. If the hypothesis is that the causal agent is either necessary or sufficient, or both, the results are simple to interpret. One definitive instance of mesothelioma (a type of cancer) occurring without any exposure to asbestos demonstrates that asbestos is not a necessary causal factor for mesothelioma. A single demonstration of a normal baby born to a mother who had taken thalidomide (a drug) shows that thalidomide is not a sufficient cause of birth defects. While these statements are true in logic, in real situations a single demonstration would not be enough, as the reliability of the observations would be questioned. While it is logically easy to *disprove* necessary or sufficient causation, it is impossible to *confirm* it; it is only possible to conclude on the basis of all reliably ascertained situations that the cause appears to be necessary or sufficient.

We need not pay much further attention to the unusual situations of necessary or sufficient causation. From now on we shall concentrate on the common situation: causation that is neither necessary nor sufficient. This causation gives a *quantitative relationship* between the causative factor and the outcome, and therefore the evidence for it must be expressed in quantitative terms.

A classification of comparative studies: classification by the groups of people compared

Cohort and intervention studies

The definition of causation leads us naturally to two basic types of study. The definition of these studies rests on the essential comparison being made. In *cohort* and *intervention* studies, groups of individuals are defined in terms of their exposure to the putative causative factor. The outcome in the ‘exposed’ group is compared with the outcome in the ‘unexposed’ group (**Ex.** 2.1). To find out if oral contraceptives cause, or protect against, heart disease, a research team could identify a group of women starting to use oral contraceptives and a group of non-users, and follow them over time to record the frequency of heart disease in each group. This is a cohort study (as it compares ‘exposed’ with ‘unexposed’ groups).



Ex. 2.1. The principle of a cohort study to test for causality. Groups of individuals are chosen, 'exposed' and 'unexposed' to the putative causal factor. The frequency of the outcome is measured in each group. The results would be expressed as:

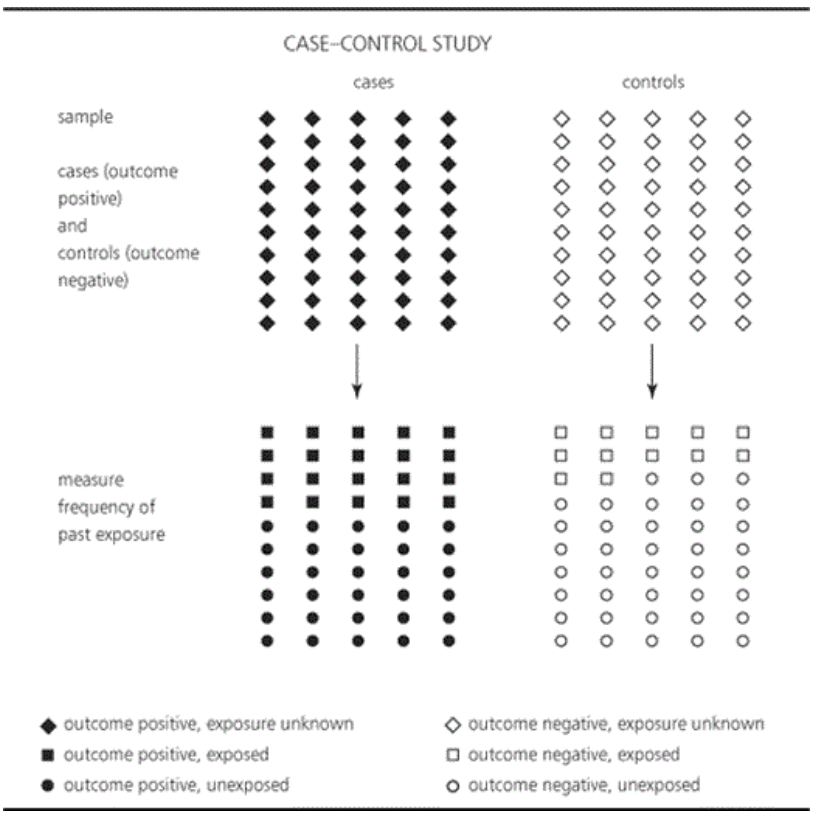
	Outcome		Total subjects
	yes	no	
Exposed group	20	80	100
Unexposed group	10	90	100

The cohort study just described is an *observational* study: the researchers observe the natural events; they do not influence which women use oral contraceptives. In *intervention* studies, the investigators control the assignment of individuals to the intervention, and so intervention studies are a special type of prospective cohort study. Thus, in the example given in Chapter 1, the clinical outcome in a group of tuberculosis patients treated with an antibiotic was compared with the outcome in a group treated without the drug.

p. 21 The term 'cohort' merely means a group of people with some common characteristic; for example, a 'birth cohort' is a group of people born in the same year or time period. Some studies merely describe the experience of such a group, and so can be called cohort studies. Thus to assess the incidence of eye disease (retinopathy) in subjects with diabetes, Younis *et al*. [1] conducted a prospective cohort study over 5 years of over 3700 patients with diabetes in England who had no retinopathy at the start of the study. As the objective was to measure the
p. 22 frequency of eye disease, no comparison group was needed. ↳ However, our interest is in comparative studies, comparing people with different exposures, who may be regarded either as subsets of a single cohort or as members of different cohorts.

Case–control studies

In *case–control* studies, groups of individuals are defined in terms of whether they have or have not already experienced the outcome under consideration, and the exposure is then measured (**Ex.** 2.2). For example, children with asthma (the 'cases') might be compared with unaffected children of the same age (the 'controls') in terms of aspects of their home environment.



Ex. 2.2. The principle of a case–control design. A group of individuals who have experienced the outcome under assessment, and a group who have not, are assessed and the frequency of exposure measured in each group. The results would be expressed as:

	Cases	Controls
	Outcome positive	**Outcome negative**
Exposed	20	12
Unexposed	30	38
Total subjects	50	50

In one of the earliest case–control studies of non-infectious disease, Dr Percy Stocks and Ms Mary Karn carried out a study in London in 1930 comparing patients with a range of types of cancer with subjects of the same age and sex without cancer, selected by the doctors of the cancer patients [2]. Many factors, such as keeping a dog or cat, having constipation, and pipe smoking, were assessed, but the conclusion emphasized by the authors was that the cancer patients ate vegetables less frequently than did the controls, suggesting a protective effect. This association was statistically significant, and the authors concluded that it 'seemed to be inexplicable from random causes' and suggested that further comparative studies, or an intervention study, could be undertaken to clarify it further. There is now substantial evidence that the intake of fresh vegetables and fruit, perhaps acting through vitamin C or other antioxidants, is protective against several types of cancer.

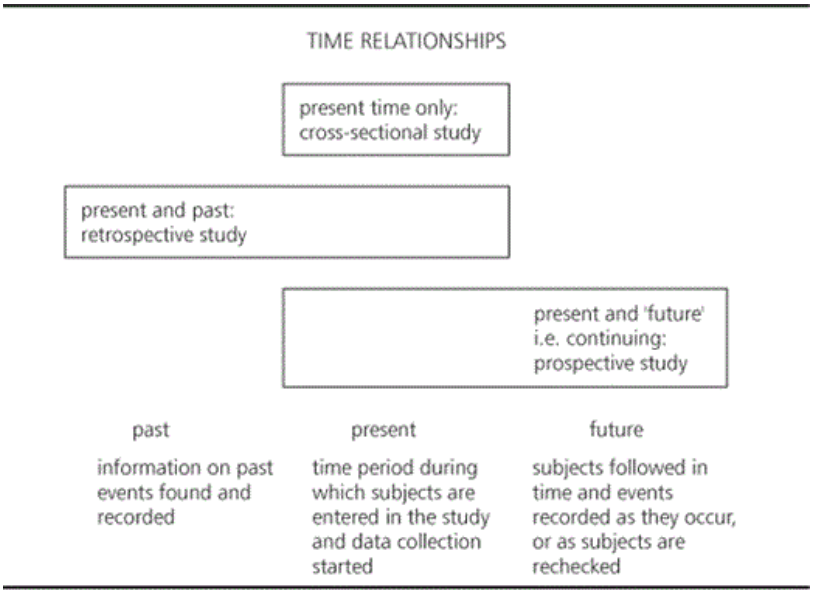
Surveys

From the sampling perspective, surveys are carried out on subjects chosen as all members or a representative sample of a population, but the sampling is not based on any defined exposure or outcome. Thus a national census, studies of all hospital admissions, or reviews of all patients in a practice or all members of a school or workforce are all surveys. As will be seen, different types of comparison may then be made within the surveyed group.

Classification by time relationships

Prospective and retrospective studies

The studies are also defined in terms of their relationship to time (**Ex.** 2.3). A study in which subjects are entered and data are collected at a point in time, and then the subjects are followed and further events recorded as they happen, is a *prospective* study. The study suggested earlier, to identify a group of women starting to use oral contraceptives and a group of non-users, and follow them ↳ over time to record the frequency of heart disease in each group, is prospective; examples of such studies will be given in Chapter 4.



Ex. 2.3. The time relationships of studies

A *retrospective* study includes observations relating to the time at which data are collected, and also previous time. Case–control studies are always retrospective, because the outcome has already happened. To explore the
p. 24 ↳ relationship between heart disease and oral contraceptive use by a retrospective case–control design, researchers would select a group of women with heart disease (cases) and a group of generally similar women without heart disease (controls), and measure for each group the frequency of exposure to oral contraceptives in the time period prior to the diagnosis of heart disease. If oral contraceptive use increases the risk of heart disease, the frequency of past use will be higher in the cases than in the controls.

Retrospective cohort studies

A cohort study can be retrospective, in that cohorts can be defined and information collected which applies to past time. Researchers may be able to identify from medical records a group of women who were using oral contraceptives 10 years ago, and a group of women who were not. They could then go to these women, or use the medical records, to determine their subsequent history from that point to the present in terms of developing heart disease. This design can be called a *retrospective cohort study*. It is retrospective because it deals with current and past events (and so may be done fairly quickly), but it is a cohort study as the comparison made is between women who used oral contraceptives and those who did not.

p. 25 In 1983 the UK Ministry of Defence commissioned a study of the health of servicemen exposed to nuclear bomb tests in Australia and the South Pacific between 1952 and 1958. In a retrospective cohort study, some 21 000 men who participated in the tests were identified and their mortality and cancer incidence from 1952 up to 1998 were compared with a comparison group of other servicemen who served in tropical and subtropical areas but were not involved in the tests. The results showed no clear increase in mortality or in any type of cancer, but there was some evidence suggesting a raised risk of some types of leukaemia [3]. Although the data were based on a follow-up of up to 45 years, the study was done in a relatively short time using the retrospective method. A major limitation is that information on other relevant exposures, including radiation exposure other than from these tests, was not available, although the choice of controls makes it likely that factors such as smoking, for example, would be similar in the two groups. Also, by relying on the Ministry of Defence records to identify both exposed and unexposed groups an estimated 15 per cent of exposed men were not identified, and it has been claimed that that omission could bias the results [4].

Cross-sectional studies

The simplest, but weakest, type of study is one where all the information is related to one point in time and is collected at that time. Thus to assess whether high blood pressure is related to blood group O, we could survey a group of subjects and determine which of them had high blood pressure, and which were of blood group O. This is a *cross-sectional* study, as the data collected relate to a cross-section in time. Where both the putative causal factor and the outcome state are stable, so that an adequate assessment can be made (e.g. blood group), this study design may be adequate. More commonly, the exposure or outcome changes over time, and so this design is weak. For example, a cross-sectional study that compares current oral contraceptive use with current high blood pressure is a very poor method of assessing an association between oral contraceptive use and high blood pressure in women. The results will be misleading, for example, if women who develop high blood pressure stop taking oral contraceptives.

To assess the relationship between housing conditions and current health, a questionnaire was sent to a 10 per cent sample of residents in one area of England [5]. Nine per cent of the respondents reported that their housing was damp, and 49 per cent of these reported long-standing illness, compared with 37 per cent of those in dry housing. The researchers analysed these data, assessing factors such as social class, employment, and various other measures. They were aware of the limitations of the cross-sectional design, and so in ↳ their discussion they considered the possibilities that the damp housing caused the ill health, but also that people with long-standing illnesses might have more difficulty maintaining their homes, and that people who are ill may for that reason gravitate to poorer housing.

p. 26

More complex designs

The feature which defines the study is the comparison which is being made: where the groups compared are defined by their exposure, the study is a cohort study; where the groups are defined by their outcome, it is a case–control study. This principle is useful when rather more complex designs are considered.

Case–control study within a cohort

In a cohort study, a large number of subjects may be enrolled while only a few suffer the outcome of interest. It may be useful to compare these subjects specifically with a sample of those without the outcome; this gives a case–control study within a cohort. For example, in a large prospective study in Japan between 1988 and 1990, over 65 000 subjects were surveyed and serum samples collected and stored. In 2002, an analysis was done comparing the 169 subjects who had subsequently developed colorectal cancer with 481 controls from the same cohort, using the data and samples collected earlier, and showed associations between serum fatty acid levels and cancer risk [6]. The study required analysis of a few hundred serum samples, rather than many thousand. As the case–control study is ‘nested’ within the larger prospective cohort study, this can be referred to as a ‘nested’ case–control study.

Cohort and case–control analyses within a cross-sectional survey

Most analyses of a cross-sectional survey will examine the frequency of characteristics and associations between them in the sampled group without considerations of causality; however, in addition to such ‘survey’ analysis, exposure or outcome based comparisons can be made. Thus a survey may be used to identify subgroups with different exposures, who are then followed giving a prospective cohort study, or subgroups can be defined in terms of an outcome and more detailed case–control comparisons can then be made.

Summary: types of study design

The different types of study are illustrated by work on a vitamin (folic acid) and a type of birth defect (neural tube defect) over some 20 years [7]. Descriptive epidemiological and clinical studies showed features
p. 27 suggesting a nutritional aetiology (such as a higher frequency in births to less affluent ↳ mothers, and time and geographical associations). A *case–control* study comparing mothers of affected babies with control mothers suggested lower levels of folic acid in the 'case' mothers [8]. Then, in an *observational prospective cohort* study, researchers measured blood levels of folic acid in about 1000 women in early pregnancy and showed that the frequency of the birth defect was higher in women with low blood levels [9]. Subsequently, in a *non-randomized intervention* study, they offered folic acid dietary supplements to some women, comparing them with others not given supplements; the frequency of the defects was reduced in those receiving supplementation, but the result was not easy to interpret as there were other differences between the two groups [10]. Finally, a *randomized intervention* study was performed, in which women were randomly assigned to receive either supplements with folic acid, or supplements without folic acid; this also showed a large reduction in the frequency of the birth defects, which is a definitive result as no other factor differed between the groups of women being compared [11]. This randomized study is described in Chapter 11.

By defining both the sampling system and the time relationships, studies can be categorized in a way that is useful in their interpretation (**Ex.** 2.4). We have now defined the most important study designs: intervention trials, observational cohort studies, case–control studies, and surveys. These main study designs have different applications, and answer different questions; the design of choice will be determined by the question to be answered, or will be dictated by practical and resource restraints. The main designs have different strengths and weaknesses, and so in assessing a study it is important to identify the design used, and then to keep these
p. 28 likely strengths and weaknesses in ↳ mind when appraising it. The method of analysis depends on the sampling system used rather than on the time relationships. This is why it is essential to understand the sampling system used in a study, as only by so doing will the correct method of analysis be chosen.

STUDY DESIGN: SAMPLING, TIME RELATIONSHIP, AND ANALYSIS

Sampling system	Time relationship	Type of analysis
COHORT: Select on exposure	prospective, retrospective, or cross-sectional	cohort
INTERVENTION: Assign exposure	prospective	cohort
CASE–CONTROL: Select on outcome	retrospective or cross-sectional	case–control
SURVEY: Total or sample of a defined population	cross-sectional or retrospective	survey, plus cohort or case–control analyses comparing subgroups

Ex. 2.4. The different types of study design

That the terminology of study design can be confusing is shown by a 1996 publication which is called a 'retrospective case–control study' in the title, but a 'historical cohort study' in the abstract [12]. The confusion appears to arise because 'case–control' is sometimes used for any study that has cases and controls, which of course can be almost any design. In this study, a case series of patients with cancer of the head and neck was used to provide two cohorts: relatives of these patients, who had the exposure of a family history of the cancer, and relatives of the spouses of the original case patients, who form the unexposed cohort. The outcome was the occurrence of head and neck cancer in these two cohorts, and so the description in the abstract was correct. In another example, the incidence of malaria was monitored in villages in Ethiopia close to small dams recently constructed to assist irrigation and was found to be seven times higher than in comparison villages that were not near dams [13]. This is a prospective cohort study, yet the accompanying editorial in the *British Medical Journal* comments on the 'weakness of the case–control design'.

Part 2. The strengths, weaknesses, and applications of the main study designs, and examples: Intervention trials

The *purpose* of an intervention trial is to measure the effect of the intervention on certain predetermined outcomes. These trials are limited (in humans) to the evaluation of interventions that are likely to be beneficial. Their main role is in the assessment of treatment, where the randomized clinical trial is accepted as the optimal method, but they are increasingly being applied to preventive programmes and to management questions as well. Intervention trials provide a direct test of the causal hypothesis that a change in the exposure factor produces a change in the outcome.

The trials are designed so that subjects receiving the intervention are similar to those in the comparison group, with regard to both any other care they receive and other characteristics that might influence the outcome being assessed. This is best achieved by using *randomization*, i.e. by selecting people who are eligible for either the intervention under test or the comparison procedure, and then selecting those who will receive the test intervention by a random process. Usually, consent is obtained before randomization, so that each subject
p. 29 needs to agree to all the possible interventions offered and to the randomization ↳ process. In large trials with an option of normal care, it may be appropriate to seek consent only from those offered the new intervention, after randomization.

The randomized intervention study has *advantages* that are shared by no other type of study. Randomization is usually the best method of obtaining two or more groups of subjects whose outcome rates, in the absence of any treatment effect, are likely to be the same. A randomized design makes it easier to use techniques such as single- or double-blind methods to minimize bias in the recording of the outcome, as will be discussed in Chapter 5. Therefore, whenever a major question arises, consideration must be given to whether a randomized intervention trial can be mounted. Often the possibility is too readily dismissed. Ineffective interventions may continue to be used, or effective interventions not widely adopted, because of the inconclusive results from less rigorous study designs. This applies not only to aspects of therapy, but also to questions of prevention, public education, and the provision of services. There is a strong case for greater use of randomized intervention studies in other social areas, such as education, management, and law.

Aside from their fundamental limitation to interventions that are likely to be beneficial, the main *disadvantages* of intervention trials relate to their requirements in terms of organization, time, cost, and resources, and in the ethical questions they raise. Clinical trials of therapy to assess modest, although important, improvements in treatment need large numbers of subjects, as will be discussed in Chapter 7. Such studies require a high degree of commitment and cooperation from health professionals and their patients. Often it is more difficult to obtain cooperation from health professionals than from patients. Large-scale trials comparing different methods for the treatment of common diseases, or assessments of preventive or screening activities, often represent triumphs of organization and persistence in the face of prejudice and apathy. Time is a problem where long follow-up is required, for example to assess survival in diseases like cancer. This may mean that the therapy assessed is obsolescent by the time the results are available. The fact that the randomized trial is an excellent scientific design does not mean that all randomized controlled trials are good or that their results are correct. A survey in 1998 of 2000 randomized trials in the treatment of schizophrenia reported that the majority were of less than 6 weeks duration, included only a small number of patients, and were of poor quality [14], and many reviews in other topics have reported similar conclusions.

Clinical trials are appropriate when a new therapy looks as if it may give better results than an established therapy. A trial will be unethical if the new therapy has very little chance of being better, or if the new therapy
p. 30 is clearly much better, making it the treatment of choice. In the 1948 trial of treatment for tuberculosis ↳ discussed in Chapter 1, some of limited supply of streptomycin was used to treat the most severe forms of tuberculosis, miliary and meningeal tuberculosis. A comparative trial was not done, as the usual outcome was so bad that if even one patient recovered after treatment with streptomycin, this would clearly be important. On the other hand, according to the UK Medical Research Council (MRC) committee, the natural history of pulmonary tuberculosis was 'variable and unpredictable' so that 'evidence of improvement or cure following the use of a new drug in a few cases cannot be accepted as proof of the effect of that drug' [15].

There is room for much professional disagreement over the necessity for a clinical trial in a particular instance; for example, there was much debate about the need for, and the ethics of, the randomized trial of folate supplementation described above. The ethical basis of involvement of randomized trials is in the uncertainty principle. It is appropriate to offer a trial to patients if the responsible clinician is uncertain as to which of the trial treatments would be best for that particular patient, and that neither the clinician nor the patient is aware of any medical or non-medical reason why one of treatments might be inappropriate for them.

Examples of intervention trials

The randomized clinical trial is accepted as the best method of investigation of new medical therapies; indeed, in many areas of therapeutics little attention is given to any studies other than randomized trials. One of the first such trials, of streptomycin treatment for tuberculosis in 1946–1948, was described in Chapter 1. It was a small quick trial (107 patients, and the main result was obtained in 6 months), which was possible because the effect size was large: a decrease in death rate from 27 to 7 per cent (**Ex. 1.8**). The *British Medical Journal* devoted its issue of 31 October 1998 to the fiftieth anniversary of this trial [16], including a reprint of the original report and several interesting papers on both past trials and issues for the future [17,18]. Sir Richard Doll gives the earliest randomized trial as the trial of immunization against whooping cough in children aged 6–18 months in the UK, which was started in 1946 but was published a few months after the report of the tuberculosis trial [17]; parents were told that half the inoculations in the study would not be active, and their consent was obtained before randomization.

The fairly late adoption of randomization in human disease trials was due to the need to gradually make this rather revolutionary method acceptable to clinicians. Randomization had been developed and used in agricultural research by R.A. Fisher and others, and had been used in trials of social programmes earlier in the
p. 31 century. Prior to this, some comparative trials used ↳ alternating treatments by order of entry or by day of admission, such as Fibiger's trial of serum treatment of diphtheria in Denmark in 1896–1887 [19], and several British studies in the 1930s. A large MRC trial in 1944 to assess the antibiotic patulin as a cure for the common cold used an alternate allocation system with two active and two placebo preparations [20]; the results showed no benefit. In a remarkable trial of a social programme to reduce delinquency, some 500 boys aged 5–13 were assigned either to a therapy programme or to a non-intervention control group by a coin toss, presumably without informed consent. A report after 30 years follow-up [21] showed worse outcomes in the intervention group in several respects. Lind's work on scurvy in 1747 used simple allocation between patients, as described in Chapter 4, and similar studies were proposed but not carried out by van Helmont in 1662.

While trials with short time frames and large anticipated effects can be quite small, detection of a modest, but important, improvement in outcome may require a very large trial. An example of a modern large-scale study is a trial assessing the effects of three different medications in the treatment of suspected myocardial infarction [22]. It involved some 58 000 patients in over 1000 hospitals in 32 countries, in a study that cost about £6 million sterling. The study showed no benefit of one widely used treatment (intravenous magnesium sulphate), despite the fact that benefits had been shown by many previous small studies. However, one of the other treatments was shown to reduce mortality significantly; treating 1000 patients for 1 month with the drug captopril would save about five lives at a drug cost per life saved of about £4000.

Trials in which two or more drugs are compared, all being given in similar ways, allow very comparable protocols and make double-blind designs relatively easy. This is more difficult where different types of care are being compared. For example, until the 1970s it was accepted that even uncomplicated suspected myocardial infarction required hospital treatment, but a trial comparing general practitioner care with immediate hospitalization showed no difference in mortality [23]. The trial required a skilled team to respond immediately when a general practitioner notified the hospital of a patient with suspected myocardial infarction and to visit the patient at home, and then randomized patients in whom both the options of hospital and home care appeared reasonable and the patient agreed to participate.

In the first trial of screening for breast cancer, some 62 000 women aged 40–64 in a health maintenance organization in New York were randomized into two equally sized groups; one group was offered the new screening intervention, mammography, and the other group received their normal care. Informed consent was obtained after randomization only from the group who were offered the extra intervention of mammographic
p. 32 screening; some 65 per cent of those ↳ women accepted the offer. No extra intervention or consent procedure was used for the comparison group, who did not necessarily know about the trial. Long-term follow-up to assess particularly the occurrence and mortality from breast cancer was instituted and is continuing. After 10–14 years of follow-up, the death rate from breast cancer in the women randomized to being offered mammographic screening was 30 per cent less than that of the comparison group [24].

The National Cancer Institute in the USA is running a complex randomized trial (the PLCO trial) to assess, in the same trial, screening techniques for prostate, lung, colorectal, and ovarian cancer, which together make up almost half of all cancers diagnosed. Over 150 000 volunteers are involved in the trial and results will accumulate over several years. Updates are available at the trial website http://www.cancer.gov/prevention/plco/.

There has been much emphasis in recent years on preserving the value of clinical trials by making sure they are done correctly and reported without bias. Standard formats for reporting results have been developed, such as the CONSORT recommended system [25], which ensure that all subjects assessed for eligibility are accounted for, whether excluded or randomized. These criteria and recommended formats are kept under review and are available from websites such as http://www.consort-statement.org. We will review this further in Chapter 4, and the CONSORT standard reporting template for a randomized trial will be shown in Ex. 4.8 (p. 96). Another emphasis has been on ensuring all trials that are performed are reported, so that all the results can be assessed to avoid publication bias. This is discussed further in Chapter 8. Investigators are encouraged to register their trial and submit the protocol when the trial is designed. Many leading medical journals will now only publish the results of trials which have been registered in advance, and which are reported according to the CONSORT criteria.

Crossover trials

Where a treatment is designed to give short-term benefit in a continuing disease, the effects of different treatments can be tested by giving them to the same patients at different times. In a randomized crossover trial, each patient is allocated each of the treatments in random order. To ensure that the effects of one treatment have completely disappeared before the other treatment is tried, a suitable 'washout period' has to be set. The analysis is a paired analysis comparing each patient's responses, which may be quantitative or qualitative, during each treatment phase compared with the other phases. There are several specific statistical issues applying to these trials, such as assessing if the time period or order of the treatments is important, and
p. 33 specific methods to ↳ estimate the appropriate study size [26]. These will not be dealt with here. In suitable applications, the design is efficient as one patient provides data on all treatments, so with two treatments the number of subjects needed is about half that of a conventional two-arm trial.

As an example, a randomized crossover study was used to assess a specific treatment (lisinopril, an angiotension-converting enzyme inhibitor) compared with placebo in the prevention of migraine [27]. Participants, who were regular migraine sufferers, were observed during a 4-week run-in period during which they were given (unknown to them) the placebo drug, and then were randomized into two groups. Thirty patients received the active drug (not knowing if it was active or placebo) for 12 weeks, discontinued it for a 2-week washout period, and were then given the placebo for a second 12-week period. The other group of 30 had the placebo drug first, and the active drug later. The patients kept a daily diary of migraine attacks and other symptoms. The trial was double blind. The main results were quantitative, assessing hours and days with headache and days with migraine, and showed on a matched analysis a statistically significant 20 per cent improvement when on the active drug. The presence or absence of adverse affects was also assessed using a paired analysis.

Individual patient (*n*-of-1) trials

In contrast with the very large studies, the efficacy of different treatments for a long-term condition in a single patient can be compared by a crossover design using randomization between different treatments given for periods of time. The '*n*-of-1' study is applicable to long-standing diseases where the objective is to improve function or decrease symptoms or disability, rather than to cure the disease. The objective is to assess the better intervention for each patient, rather than the one which is better on average; but a series of such studies can assess treatments more generally [28]. An example is a comparison of two different medications for osteoarthritis, assessing 2-week courses of each of the treatments in randomized order [29]. Other uses have been in hypertension and chronic respiratory disease. A small survey in primary care showed that patients participating in these trials appreciated their active involvement, for example through keeping symptom diaries, and had a sense of empowerment and control [30]. Mahon *et al*. [31] used a conventional two-arm randomized trial to compare conventional care with the option of selecting medication by *n*-of-1 trials. In this study, patients with chronic respiratory disease were randomized either to have an *n*-of-1 trial comparing a drug (theophylline) against placebo, continuing with the choice which gave a better outcome on the basis of the *n*-of-1 trial, or to have standard ↳ management where the theophylline was continued if the patient's respiratory disease had improved while using it. While the preliminary results showed a benefit in lower drug use at 6 months follow-up, later results showed no differences at 1 year follow-up on larger numbers of patients.

Community-based cluster trials

Where the intervention is made on a community basis, it is appropriate to assign the intervention and make the comparisons amongst groups of people, from families to communities. Thus in an intervention study in Ethiopia, households were provided with either iron cooking pots or aluminium pots on a random basis, and the change in haemoglobin level in one child per household assessed over 12 months, with almost all children being followed; this showed that children in households supplied with iron pots had lower risks of anaemia and improved growth [32].

To evaluate the impact of improved water supplies or village health workers' activities in a developing country, comparisons between villages are appropriate. To evaluate health promotion activities, comparisons between towns, workforces, or schools may be appropriate; for example, studies of the prevention of coronary heart disease have compared workforces receiving a complex intervention programme with those not receiving it. Differences in the calculation of the necessary size of the study, and many other methodological issues, are involved when communities rather than individuals are compared [33].

Health care intervention may be done in general practices or hospitals. A trial of a new method for the diagnosis of skin lesions in primary care, with the objective of reducing unnecessary excisions, involved randomizing 223 general practices in Perth, Australia, equipping and training the 468 doctors in the intervention practices to use the new method, and comparing the ratios of benign to malignant lesions excised in the intervention and the control practices over a year [34]. The design avoided the consent processes needed if patients were to be individually randomized, and was relevant as the new system would be implemented at a practice level; but the results showed no benefit of the new method.

Non-randomized intervention trials

Because of the logistical or ethical problems of randomization, non-randomized interventional studies may be used. However, without the strengths that randomization gives in avoiding biases, non-randomized intervention trials are more difficult to interpret. Indeed, non-randomized intervention studies may be best considered as observational cohort studies. They are especially valuable in situations where randomization would be very difficult or unacceptable.

p. 35 Such intervention trials include studies of fluoridation of water supplies in which communities using fluoridated water were compared with generally similar communities continuing with unfluoridated supplies [35]. In health education, the Stanford Heart Prevention Project compared intensive methods of health education in two communities, comparing them with rather similar communities that did not receive the intervention [36,37]. A recent non-randomized study of screening has had very clear results. Neuroblastoma is a cancer affecting young children, which can be detected by a urine test in young children, and screening has been advocated, especially in Japan. In Canada, a comparison was made between a cohort of births in Quebec, where a screening programme was set up, and a comparison cohort, unscreened, in Ontario. The results showed that screening produced a great increase in the incidence rate of diagnosed neuroblastoma, but the mortality rate was not reduced. The interpretation was that screening produced a substantial number of false-positive diagnoses, and also 'silent tumours', which are tumours which were diagnosed as neuroblastoma and treated, but which if left alone would not have produced any clinical problems [38]. The results prevented further programmes from being set up, and led to the discontinuation of neuroblastoma screening in Japan.

In clinical trials of therapies, the outcome can be easily influenced by differences in the subjects selected for different treatments; the numbers of subjects are usually small, so that non-randomized trials are regarded with considerable suspicion. There are many examples of situations in which apparently similar groups of patients treated by apparently similar methods give considerably different outcomes, with no ready explanation. Several examples in surgery, such as gastric freezing, were documented in an influential book [39]. This treatment for peptic ulcer, using a gastric balloon with a coolant, was widely used in the USA in the 1960s on the basis of a logical argument that cooling reduces gastric acid secretion, and the unsystematic comparison of results after the treatment with past experience. A randomized trial then showed no benefit, and indeed somewhat worse results, when the treatment was compared with a sham treatment [40].

An unfortunate example is a study to compare the survival of patients with breast cancer attending a centre for alternative medicine in England with a comparison group selected from patients attending conventional hospitals [41]. The great difficulty in this type of study is to ensure comparability between the two groups in all other factors that relate to survival, particularly the extent (size and staging) of the breast cancer before treatment. This study showed worse survival in patients seen at the alternative medicine centre. It created
p. 36 much controversy, and subsequently the authors stated that they did ↳ not think the reduction in survival was due to the alternative medicine regimen, but was more likely to be due to a greater initial severity of disease in those who attended the centre. However, this explanation was not given prominence in the original paper, which was regarded by many as being seriously in error. Many correspondents felt that the problems in interpretation showed that only randomized trials should be used to assess different modes of treatment. Indeed, a randomized trial had been proposed, but agreement for such a trial could not be obtained from all parties.

A still common, but unsatisfactory, design is to compare the results of patients given a new treatment with previous patients given the previously used treatment, sometimes called *historical controls*. This is a particularly dangerous comparison, as there are many factors that may influence why a particular patient is offered or accepts the new treatment. A slightly better comparison may be to compare all patients undergoing treatment in a particular facility after the new treatment is introduced with all patients seen at the facility before it was introduced, thus avoiding the problems of individual selection and acceptance of the new treatment. However, this design is dependent on there being no change in the factors that influence patients to go to that particular facility (i.e. the referral patterns). Also, when a new intervention is introduced, the professionals who deliver it may well act differently to those who continue to use the conventional programmes. A new educational approach in schools, for example, is very likely to be first taken up and promoted by particularly interested teachers, and it may be the differences between these teachers and others which affect the educational outcome, rather than particular properties of the new programme.

In contrast, other non-randomized trials have given results that have been confirmed by a randomized study. In a non-randomized study of the prevention of congenital defects by dietary supplementation, women who came for genetic counselling in good time for their next pregnancy were offered and accepted the dietary supplement. They were compared with women who did not receive the supplement, in most cases because they did not seek professional help early enough in regard to their next pregnancy. The comparison showed a considerable reduction in birth defects in women receiving the supplement, but this could obviously be influenced by a large number of factors which would be related to whether the woman sought professional help in time for her to be offered supplementation [10]. However, a further study showed that in the area of England where this study was carried out, the recurrence rate of the defect in *all* births to at-risk women fell over time, at the same time as the proportion of all at-risk women receiving this dietary supplementation rose [42]. While
p. 37 still a non-randomized comparison, and open to the ↳ major problem that something else might be changing over time which affects the recurrence rate, this second comparison does avoid the individual selection effects which complicate the comparison between women being offered and not being offered the supplement. As noted earlier, these results were later confirmed by a randomized trial, which reported in 1991 [11]; it is an interesting question whether more action should have been taken after the 1981 results, rather than waiting 10 years for the later results.

Observational cohort studies

The other types of studies involve observation rather than intervention, i.e. the researcher does not control the intervention, but merely observes its effects. A cohort study is designed to answer the question: 'What are the effects of this particular exposure?' The studies compare a group of people with the exposure under consideration with a group without the exposure, or with a different level of exposure. In a prospective cohort study, the exposure is defined and the subjects are selected before the outcome events occur. This ensures that the time relationship between exposure and outcome is appropriate; the follow-up procedures allow direct measurement of the incidence rate of the outcome(s) in each of the groups being studied.

Therefore cohort studies are the appropriate method for studying the effects of a certain exposure. Thus cohort studies were set up to follow survivors of the atomic bomb explosions in Japan in 1945 and those affected by the Chernobyl nuclear plant accident in 1986, assessing many endpoints, for example total mortality or the incidence of leukaemia. The effects of many occupational hazards have been established by studying cohorts of exposed workers, comparing their subsequent mortality or morbidity either with defined comparison groups or with the general population, which can be regarded as unexposed to the particular chemical or process.

Several studies before the Second World War showed lung scarring, and suggested a possible increase in lung cancer, in workers exposed to asbestos. Increases in several types of cancer were then shown in cohort studies, first in the UK and then in the USA [43]. These studies led to much stricter control of asbestos exposure in workplaces and in the general population, and also to legal actions that forced a major asbestos manufacturer into bankruptcy.

A major disadvantage of prospective cohort studies relates to the time required. If the outcome does not occur until after a long period the exposure, the study has to have many years of follow-up. We shall discuss in Chapter 4 the design of studies in which women using oral contraceptives were recruited either through particular clinics or through their regular doctors, and followed up over time to assess many different types of
p. 38 morbidity and mortality. ↳ A retrospective cohort design will overcome this problem, but is possible only if there are adequate records for the appropriate subjects and the time period.

If the outcome does not occur frequently, the cohort study must involve large numbers of subjects to observe even a reasonable number of events. Issues of the necessary size of the study will be discussed in Chapter 7. The effective size of a study depends primarily on the number of outcome events that are observed. For example, in a prospective study of pregnancies to assess the association between a drug for maternal epilepsy and congenital malformations, data from 2332 pregnancies were obtained, but the critical figure is the number of malformed births occurring to exposed mothers, which was only six [44]. A later and larger study of this topic is discussed in the next chapter.

Large cohort studies

The Million Women Study was set up in the UK to study the relationship between hormone replacement therapy and breast cancer incidence and mortality. Between 1996 and 2001, women coming for breast cancer screening were invited to join the study and to provide important baseline information. Follow-up was by linkage within the NHS central register to cancer registrations and deaths. As of 2003, over one million women had joined the study, with an average age of 56, and over 50 per cent had used hormone replacement therapy at some time. Over 9000 breast cancers have been reported in these women [45]. The study has enormous potential to assess other exposure factors and other disease outcomes. The study website (http://www.icnet.uk/research/studies/mws/index2.html) provides current updates and lists of publications.

The Million Women Study emphasizes relatively simple data collection at one point in time and routine follow-up through data linkage to achieve a very high sample size. This contrasts with the European Prospective Investigation into Cancer and Nutrition (EPIC), which started in 1992 and includes over 500 000 participants in 10 European countries. The participants were asked to provide extensive data including food consumption questionnaires, lifestyle data, and blood samples, resulting in the collection of nine million specimens, the largest repository of biological material in a study in the world [46]. Participants have been followed up regularly by multiple methods, and by 2004 over 20 000 cancer cases had been reported. Many of the analyses will be in the form of nested case–control studies; again a useful website gives current information (http://www.iarc.fr/epic).

Subjects may be chosen for cohort studies not because they are particularly unusual in terms of exposure, but because they can be kept under surveillance over a long period. An example of this type of study is the cohort
p. 39 study of the ↳ effects of smoking in British doctors started in the 1950s. Doctors were selected not because their smoking habits were particularly unusual, but because they could be followed through annual registration procedures and their rate of cooperation was expected to be high [47,48]. Using similar logic, a cohort study of American nurses was started in the late 1970s as an opportunity to assess many major health issues in women [49,50].

A cohort can be set up by doing an initial survey of a sample of subjects or of a whole community, chosen not because of any unusual characteristic of exposure but because of availability and practical considerations. Such general population follow-up studies are analysed as cohort studies comparing subjects with different exposures from within the same initial group. An important example is the Framingham Heart Study, started in the 1950s by asking for volunteers from the small town of Framingham, a short distance from Boston, Massachusetts, enrolling those who volunteered, and following them up over a long period of time [51,52].

Cohort studies from early life

Several cohort studies have started at birth or early childhood with active follow-up. A study based on a sample of 4454 people born during a single week, 3–9 March 1946, in England, Scotland, and Wales has followed them through childhood and adolescence into adulthood; active follow-up is continuing [53,54]. A similar study in New Zealand of a group of children identified at birth in 1972–1973 in a single town has continued by inviting participants to come for a one or two day assessment of psychological and physical characteristics about every 3 years, and is now continuing to assess the children born to the original cohort members. A study of asthma from this cohort is assessed in detail in Chapter 12. An ambitious proposal for a 'conception to death' cohort study is aimed at addressing prenatal influences and events [55]. A plan for a large prospective cohort study of 500 000 individuals aged 45–69 puts emphasis on assessing biochemical and genetic parameters, using an initial questionnaire and blood sample with analyses based on different biological parameters being done later; this UK Biobank project has created considerable controversy [56].

Case–control studies

The question asked in a case–control study is: 'What are the factors which caused this event?' Case–control studies start after the outcome event, and therefore have the opportunity to assess multiple causes relating to one event. Thus they are the prime method of assessing the causes of a new problem. A good example of the investigation of a newly recognized problem is given by ↳ studies of cancer of the vagina occurring in young women [57]. This disease had been generally described as rare, and occurring in older women, but between 1966 and 1969 seven occurrences in women aged 15–22 were reported in Boston. To investigate possible causes, these cases, plus one further patient seen in another Boston hospital, were studied, and for each 'case', four control subjects were chosen, females born within 5 days of and in the same hospital as the case women. A detailed interview schedule was drawn up, and interviews were carried out in an identical fashion on all 40 subjects. These interviews explored a wide range of possible factors, and the greatest difference between the cases and controls was in the use of diethylstilboestrol (DES) by their mothers during the pregnancy, which was documented for seven of the eight cases, but for none of the 32 controls. This suggested that it was the key causal agent, a result confirmed by further studies.

p. 40

The case–control study is a natural development from the case series. If we identify people with a condition of unknown cause, we start looking for possible causes by considering the characteristics of these subjects, and we then have to consider how common those characteristics would be in an unaffected population. Beyond this, the major *advantage* of case–control studies is that they are highly efficient in terms of the numbers of subjects required. Thus, for example, in a study of the causes of breast cancer, in principle the amount of information given by studying 100 women who have developed breast cancer and 100 controls is similar to that produced by asking the same questions of a group of some 50 000 middle-aged women and following them for a year, during which 100 of them might develop breast cancer.

This advantage of numbers is because the outcome has already happened. However, that fact also results in the great *disadvantage* of case–control studies, that all information on previous events has to be collected after the outcome has happened. This produces two major problems. First, the relevant causal events may have happened a considerable time in the past, and therefore the information obtained is likely to be incomplete and inaccurate. However, the greater problem is that the case subjects, who have suffered the outcome under investigation, may tend to give different responses to the control subjects, thus introducing a bias into the information obtained on previous exposures. These problems, and methods to overcome them, will be discussed further in Chapter 4.

The advantages in the numbers of subjects and the time required for a case–control study must be compared with the disadvantages of unreliability of exposure information and possible bias in the recall of events. Often these biases are severe enough to justify a more expensive and time-consuming cohort study. Case–control studies have been of immense importance in elucidating the causes of chronic diseases, such as cancer, and the
p. 41 first indications of ↳ causes have often come from such studies. A good example is the association between smoking and lung cancer.

Between 1922 and 1947, the number of deaths from lung cancer in England and Wales increased by a factor of 15. While some thought this might be an artefact of diagnosis or recording, investigations suggested that the increase was real; the most favoured explanations were an effect of atmospheric pollution or of smoking tobacco, and small studies were available on each. In 1947, the MRC decided to carry out a more thorough investigation. To do this, patients with lung cancer and some other cancers in some 20 London hospitals were interviewed, along with control patients who had a range of diseases excluding cancer. The results showed that 647 of 649 men with lung cancer were smokers (99.7 per cent), but 622 out of 649 (95.8 per cent) of the control patients were also smokers. The results by detailed smoking history were more useful, and showed that the risk of lung cancer increased strongly with the amount smoked [58]. The study was then expanded, with similar results [59]. The original report of this investigation is well worth reading because of the detail given to the methodology and the consideration of biases and possible explanations of the results. For example, the issue of possible bias by the subject or the interviewer is addressed by looking at the results from a number of patients who, at the time of interview, were thought to have lung cancer, but in whom the final diagnosis was something else. The smoking habits of these patients were clearly different from those of patients who actually had lung cancer, but were similar to those of the non-cancer control patients, which would not have happened if the excess smoking recorded for lung cancer patients had been due to bias in the observations.

This was one of the first case–control studies, and acceptance of its results required several years of debate about, amongst other things, the value of the case–control approach. This study produced results in 1950; the prospective study of British doctors, published 14 years later, gave substantially the same results [47,60], as shown in **Ex** 2.5.

RISKS GIVEN BY DIFFERENT STUDIES

Relative risks of lung cancer by daily cigarette smoking from different studies

Non-smokers (referent)	1.0	1.0	1.0	1.0
1–14	11.0	7.0	8.1	7.8
15–24	13.9	9.5	19.9	12.7
25+	27.0	16.3	32.4	25.1
Publication date:	1950 (1)	1952 (2)	1964 (3)	1976 (4)

1. Doll and Hill (1950): 3 years' study: 1298 subjects: case–control, incidence, odds ratio [58].
2. Doll and Hill (1952): 5 years' study: 2714 subjects: case–control, incidence, odds ratio [59].
3. Doll and Hill (1964): 13 years' study: 34 445 subjects: cohort, mortality, relative risk [47].
4. Doll and Peto (1976): 25 years' study: 34 445 subjects: cohort, mortality, relative risk [60].

Ex. 2.5. Comparison of results. The relationship between smoking and lung cancer in men, as shown by different British reports. Relative risk, or its close approximation odds ratio, is the basic measure of association given by these studies and shows the ratio of the frequency of lung cancer at each level of smoking compared with the frequency in non-smokers (see Chapter 3)

Applications of case–control studies

Although their main contribution has been to aetiological studies, case–control studies are applicable to many other questions. For example, the effect of population-based interventions can be assessed; studies of screening for cancer have been done by comparing subjects who die from the diseases in question with samples of the unaffected population [61,62]. The benefits of screening are shown by a less frequent history of screening in those who suffer the outcome which screening is designed to prevent, such as death or advanced
p. 42 disease; although the interpretation of these studies can be difficult [63]. ↳ For example, no randomized trial has yet been done to assess whether routine examination of the skin results in ultimate benefit, i.e. a reduction in mortality from the most serious skin cancer, melanoma. However a case–control study in which information was gathered from melanoma cases and comparison subjects at the time of diagnosis, and then analysed as a comparison between the skin cancer cases who went on to develop advanced disease and the control group, gives results suggestive of some benefit [64].

Case–control studies deserve to be more widely used in assessments of such population interventions, and also in assessments of different types of medical care. An interesting example of the case–control approach to a clinical topic is a study of prognostic factors in breast cancer in which long-term survivors were compared with patients who had died [65]. The method is also used in non-health areas, although often with different terminology. For example, to assess the risk of bankruptcy, studies have compared companies that went bankrupt with a sample of companies that did not, in terms of various financial parameters. These were case–control studies, and the interpretation of their results involves the same issues as apply to health-related studies [66].

Case–control studies can be quite large. In an expansion of the routine function of disease registries in Canada, information has been collected routinely from newly diagnosed cancer patients and controls chosen from the
p. 43 population ↳ with a similar age and sex distribution, to give a database which can be analysed for many exposure factors and different types of cancer. An analysis has been done in regard to obesity, using data on 21 000 cancer cases and 5000 population controls. It found a significant increase in cancer risk overall and increased risks for several individual cancers, and concluded that excess body mass accounted for 7.7 per cent of all cancers in Canada [67].

Unfortunately, there are many examples of badly performed case–control studies, which have led to erroneous results. There are more examples of poor quality case–control studies than there are of poor cohort studies. This is partly because the case–control design is more subtle and more difficult to apply than a cohort design, but also because case–control studies may appear deceptively easy. There has been a simplistic view that case–control studies are 'quick, simple, and cheap' compared with cohort studies, whereas in fact they may not be; to mount a well-designed case–control study of an important question is often a complex and expensive endeavour. However, because of this impression, many case–control studies have been done with inadequate resources and poor design, and not surprisingly have given erratic results. The issues are those of appropriate study design and quality control of the information collected, which will be discussed in Chapters 4 and 5, and of appropriate analysis, sample size, power, and publication bias, which will be discussed in Chapters 6 and 7.

Case–crossover aetiological studies

The equivalent of the crossover trial in an aetiological study is to look at factors related to an episodic event, comparing them with a similar time period for the same individuals that did not result in an event. This is often called a case–crossover design. Thus in a study to assess the role of mobile phone use in motor vehicle accidents, drivers who had reported a motor vehicle accident were identified, and 74 699 of these drivers who owned a mobile phone allowed access to their phone records [68]. The study was done in Toronto before there were any legal restrictions on mobile phone use while driving. The analysis consisted of assessing from these records whether the mobile phone was being used immediately before the reported accident compared with use by the same driver in a set of control periods defined, for example, as the same time on the day before, the same time and the same day of the week before, and so on. The analysis showed an association: an increased relative risk of a motor vehicle accident if the mobile phone call was between 1 and 5 minutes prior to the accident, with the risk decreasing as the time interval between the call and the accident increased, up to 16–20 minutes. The relative risk was 4.8, and was no lower where a hands-free mobile phone was used.

p. 44

Cross-sectional surveys

The basic purpose of a survey is to measure the prevalence (frequency at one point in time) of a condition. It is therefore the appropriate method to answer questions such as: Is anaemia common in this community? What is the prevalence of hypertension in elderly people? How satisfied are patients with the health services? Surveys also allow the assessment of associations, such as whether hypertension or a poor diet is more common in one social group than another, and such comparisons can lend themselves to causal thinking. Thus Pauletto *et al*. [69] surveyed two groups of villagers in Tanzania and showed that the group who consumed a fish-based diet had lower blood pressure than the group who had a vegetarian diet. Cross-sectional surveys are limited to making observations that apply to one point in time, and therefore judgements about causation based on survey data have to be very cautious. This is because causation by its definition includes a time function, and therefore, to assess it, studies must cover a time period either prospectively or retrospectively. Our definition of a cross-sectional survey is quite narrow; if a population is surveyed to find out which subjects have a poor diet, and these subjects are then compared with those with a better diet in terms of subsequent health care utilization, we classify the first part as a survey, but the second as a prospective cohort study. Thus, a health survey was carried out in two areas of Sweden between 1963 and 1965, and subsequent linkage of the data to mortality and cancer registration data up to 1987 allowed a prospective cohort analysis of the association between obesity and cancer risk [70].

The advantages of a cross-sectional survey arise primarily because of its relative simplicity, as the complications of retrospective data collection and prospective follow-up are both avoided. Thus the methods to be used can be extensively pre-tested, applied on a large scale, and made reproducible, and at least single blind, in that the investigators need know nothing about the state of the individuals being assessed. Because of their simplicity, cooperation can be high as little is demanded of the subjects.

As the primary function of a survey is to assess prevalence, the methods obtained to draw samples from a defined population are crucial, and various randomized or systematic methods are available which have been discussed in detail in standard texts [71–73]. The use of such methods can ensure that the subjects surveyed are a representative sample of the community, and allows the repetition of the survey at other times or in other communities using identical methods and yielding comparable results. Thus surveys can monitor changes in a population in time and assess differences between population subgroups.

p. 45

The main disadvantages of survey methods also arise from their simplicity. Because of their lack of a time dimension, the interpretation of their results in terms of cause and effect is very limited, and over-interpretation in this regard is a considerable danger. Their power in assessing prevalence, and therefore in assessing associations between different states, requires a reasonable prevalence of that condition in the population, otherwise they will be inefficient.

Summary of comparisons

The major properties of the different study designs are summarized in **Ex.** 2.6. In deciding which approach to adopt to deal with a specific question, these properties can be used as a general guide, but the final decision on the appropriate method will be based on a detailed consideration of the issues to be addressed in the study, including the likely effects of bias, confounding, and chance on the interpretation of the results. These concepts will be explored in subsequent chapters.

MAIN PROPERTIES OF DIFFERENT STUDY DESIGNS

Design	Intervention trials	Cohort studies	Case–control studies	Surveys
Question asked	What is the effect of this intervention?	What are the effects of this exposure?	What were the causes of this event?	How common is this condition? Are conditions and exposures associated?
Applicability	Controlled interventions of likely benefit	Any exposure for which adequate numbers of exposed subjects can be found and studied, and outcome can be assessed	Any event for which groups of cases and appropriate controls can be found, and exposure factors can be assessed retrospectively	Any exposure, condition, or association which is reasonably common and for which assessment at one point in time is sufficient
	Primary method of studying new therapies	Primary method of studying unusual or new exposures	Primary method of studying unusual or new outcomes	Primary method of assessing prevalence
Major strengths	Allows randomization: best way to control confounding Allows double-blind assessment: best way to control bias	Allows multiple endpoints to be assessed Cause to effect time sequence clear All measures of risk can be assessed Exposure is assessed prior to outcome, avoiding bias	Usually can be done with moderate numbers of subjects; feasible even on small numbers Retrospective method is rapid Multiple exposure factors can be assessed	Representative samples of a population can be drawn Methods can be standardized, reliable, and single blind Efficient in resources needed Cooperation may be high Can be repeated using similar methods
Major weaknesses	Ethical limitations Organizational problems Time scale	Usually requires large numbers Long time scale for some effects	Retrospective method limits exposure information and is open to bias Adequate control group may be difficult to define or obtain	Lack of time dimension limits causal interpretations Inefficient for rare exposures or conditions

Ex. 2.6. Some properties of the four major study designs

Self-test questions (answers on p. 492)

Q2.1 What types of study design are illustrated by the following situations? (Consider the groups being compared, and the time relationship).

(a) The frequency of deep venous thrombosis in people who have recently had a long-distance air flight is compared with the frequency in an age-matched sample of people who have not travelled.

(b) To assess a possible protective effect of exercise, past involvement in competitive sports is assessed for women who have hypertension and for a group of women of the same age who have not.

(c) Total mortality over the last 30 years is assessed in men employed in a uranium mine at some time during that period, and compared with the general population.

(d) To assess an effect of smoking, men admitted to hospital with myocardial infarction are compared with similar unaffected men in the community.

(e) In a continuation of the same study, the risk of future repetition of the heart attack (reinfarction) is compared between those who smoked and those who did not at the time of their first heart attack.

(f) Groups of breast-fed and non-breast fed babies are identified at 6 months of age, and then surveyed at later times. To study the association with small stature, children in the lowest decile
p. 46 of height at ↳
p. 47 ↳
age 10 years are compared with a sample of all the other children in regard to many factors.

(g) A dermatologist offers a new treatment for acne to all her teenage patients, and compares the results with those who chose to continue the traditional treatment.

Q2.2 Do you agree with the following conclusions?

(a) A study shows higher rates of diabetes in unemployed than in employed men. This shows that becoming unemployed leads to diabetes.

(b) A survey of residents in an aged care facility shows that smoking is much less frequent in 95-year-olds than it is in 85-year-olds. This shows that many people give up smoking between these ages.

(c) In a study of breast cancer patients, 40 per cent reported stressful life events in the last 2 years compared with only 25 per cent of age-matched controls. This shows that stress increases the risk of breast cancer substantially.

(d) A survey of workers in a clothing factory shows that the frequency of repetitive strain injury is lower than the rate reported in a general population survey. This shows that repetitive strain injury is not caused by this type of work.

Q2.3 What type of study is most useful to assess the various consequences of an environmental hazard, such as radiation exposure? What are its main disadvantages, in terms of an outcome such as cancer occurrence?

Q2.4 What are the prime advantages and disadvantages of a case–control study in assessing the causes of a newly recognized disease?

Q2.5 What are the main advantages of the randomized trial design?

References

1. Younis N, Broadbent DM, Vora JP, Harding SP. Incidence of sight-threatening retinopathy in patients with type 2 diabetes in the Liverpool Diabetic Eye Study: a cohort study. *Lancet* 2003; **361**: 195–200. 10.1016/S0140-6736(03)12267-2
WorldCat Crossref

2. Stocks P, Karn MN. A co-operative study of the habits, home life, dietary and family histories of 450 cancer patients and of an equal number of control patients. *Ann Eugen* 1933; **5**: 237–280. 10.1111/j.1469-1809.1933.tb02094.x
WorldCat Crossref

3. Muirhead CR, Bingham D, Haylock RGE, O'Hagan JA, Goodill AA, Berridge GL et al. Follow up of mortality and incidence of cancer 1952–98 in men from the UK who participated in the UK's atmospheric nuclear weapon tests and experimental programmes. *Occup Environ Med* 2003; **60**(3): 165–172. 10.1136/oem.60.3.165
WorldCat Crossref

4. Baverstock K. The 2003 NRPB report on UK nuclear-test veterans. *Lancet* 2003; **361**(9371): 1759–1760. 10.1016/S0140-6736(03)13447-2
WorldCat Crossref

p. 48 5. Packer CN, Stewart-Brown S, Fowle SE. Damp housing and adult health: results from a lifestyle study in Worcester, England. *J Epidemiol Community Health* 1994; **48**: 555–559. 10.1136/jech.48.6.555
WorldCat Crossref

6. Kojima M, Wakai K, Tokudome S, *et al*. Serum levels of polyunsaturated fatty acids and risk of colorectal cancer: a prospective study. *Am J Epidemiol* 2005; **161**(5): 462–471. 10.1093/aje/kwi066
WorldCat Crossref

7. Elwood JM, Little J, Elwood JH. *Epidemiology and Control of Neural Tube Defects*. Oxford: Oxford University Press, 1992.
Google Scholar Google Preview WorldCat COPAC

8. Hibbard ED, Smithells RW. Folic acid metabolism and human embryopathy. *Lancet* 1965; **1**: 1254–1256. 10.1016/S0140-6736(65)91895-7
WorldCat Crossref

9. Smithells RW, Sheppard S, Schorah CJ. Vitamin deficiencies and neural tube defects. *Arch Dis Child* 1976; **51**: 944–950. 10.1136/adc.51.12.944
WorldCat Crossref

10. Smithells RW, Sheppard S, Schorah CJ, *et al*. Apparent prevention of neural tube defects by periconceptional vitamin supplementation. *Arch Dis Child* 1981; **56**: 911–918. 10.1136/adc.56.12.911
WorldCat Crossref

11. MRC (Medical Research Council) Vitamin Study Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. *Lancet* 1991; **338**: 131–137. 10.1016/0140-6736(91)90133-A
WorldCat Crossref

12. Foulkes WD, Brunet J-S, Sieh W, Black MJ, Shenouda G, Narod SA. Familial risks of squamous cell carcinoma of the head and neck: retrospective case–control study. *BMJ* 1996; **313**: 716–721. 10.1136/bmj.313.7059.716
WorldCat Crossref

13. Ghebreyesus TA, Haile M, Witten KH, *et al*. Incidence of malaria among children living near dams in northern Ethiopia: community based incidence survey. *BMJ* 1999; **319**: 663–666. 10.1136/bmj.319.7211.663
WorldCat Crossref

14. Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50 years. *BMJ* 1998; **317**: 1181–1184. 10.1136/bmj.317.7167.1181
WorldCat Crossref

15. Silverman WA. *Where's the Evidence? Debates in Modern Medicine*. Oxford: Oxford University Press, 1998.
Google Scholar Google Preview WorldCat COPAC

16. British Medical Journal. Fifty years of randomised controlled trials. *BMJ* 1998; **317**: 1167–1248.
WorldCat

17. Doll R. Controlled trials: the 1948 watershed. *BMJ* 1998; **317**: 1217–1220. 10.1136/bmj.317.7167.1217
WorldCat Crossref

18. Yoshioka A. Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *BMJ* 1998; **317**: 1220–1223. 10.1136/bmj.317.7167.1220
WorldCat Crossref

19. Fibiger J. Om Serumbehandlung af Difteri. *Hospitalstidende* 1898; **6**: 309–325.
WorldCat

20. Chalmers I, Clarke M. Commentary: The 1944 patulin trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council. *Int J Epidemiol* 2004; **33**(2): 253–260. 10.1093/ije/dyh162
WorldCat Crossref

21. McCord J. A thirty year follow-up of treatment effects. *Am Psychol* 1978; **33**: 284–289. 10.1037/0003-066X.33.3.284
WorldCat Crossref

22. ISIS-4 (Fourth International Study of Infarct Survival Collaborative Group). ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet* 1995; **345**: 669–685. 10.1016/S0140-6736(95)90865-X
WorldCat Crossref

23. Hill JD, Hampton JR, Mitchell JRA. A randomised trial of home-versus-hospital management for patients with suspected myocardial infarction. *Lancet* 1978; **i**: 837–841. 10.1016/S0140-6736(78)90190-3
WorldCat Crossref

24. 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

25. 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

p. 49

26. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol* 1979; **8**: 7–20. 10.1111/j.1365-2125.1979.tb05903.x
WorldCat Crossref

27. Schrader H, Stovner LJ, Helde G, Sand T, Bovim G. Prophylactic treatment of migraine with angiotensin converting enzyme inhibitor (lisinopril): randomised, placebo controlled, crossover study. *BMJ* 2001; **322**: 19–22. 10.1136/bmj.322.7277.19
WorldCat Crossref

28. Cook DJ. Randomized trials in single subjects: the *N* of 1 study. *Psychopharmacol Bull* 1996; **32**: 363–367.

29. March L, Irwig L, Schwarz J, Simpson J, Chock C, Brooks P. *n* of 1 trials comparing a non-steroidal anti-inflammatory drug

with paracetamol in osteoarthritis. *BMJ* 1994; **309**: 1041–1046. 10.1136/bmj.309.6961.1041
Crossref

30. Nikles CJ, Clavarino AM, Del Mar CB. Using *n*-of-1 trials as a clinical tool to improve prescribing. *Br J Gen Pract* 2005; **55**: 175–180.

31. Mahon JL, Laupacis A, Hodder RV, *et al*. Theophylline for irreversible chronic airflow limitation: a randomized study comparing *n* of 1 trials to standard practice. *Chest* 1999; **115**: 38–48. 10.1378/chest.115.1.38
Crossref

32. 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

33. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ, Donner A. Methods in health service research. Evaluation of health interventions at area and organisation level. *BMJ* 1999; **319**: 376–379. 10.1136/bmj.319.7206.376
WorldCat Crossref

34. 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

35. Arnold FA, Dean HT, Jay P, Knutson JW. Effect of fluoridated public water supplies on dental caries prevalence. Tenth year of the Grand Rapids–Muskegon study. *Public Health Rep* 1956; **71**: 652–658. 10.2307/4589490
WorldCat Crossref

36. Farquhar JW, Fortmann SP, Maccoby N, *et al*. The Stanford Five-City Project: design and methods. *Am J Epidemiol* 1985; **122**: 323–334.
WorldCat

37. Fortmann SP, Flora JA, Winkleby MA, Schooler C, Taylor CB, Farquhar JW. Community intervention trials: reflections on the Stanford Five-City Project experience. *Am J Epidemiol* 1995; **142**: 576–586.
WorldCat

38. Soderstrom L, Woods WG, Bernstein M, Robison LL, Tuchman M, Lemieux B. Health and economic benefits of well-designed evaluations: some lessons from evaluating neuroblastoma screening. *J Natl Cancer Inst* 2005; **97**: 1118–1124. 10.1093/jnci/dji203
WorldCat Crossref

39. Bunker CB, Barnes JA, Mosteller F, (Eds.). *Costs, Risks and Benefits of Surgery*. New York: Oxford University Press, 1977.
Google Scholar Google Preview WorldCat COPAC

40. Ruffin JM, Grizzle JE, Hightower NC, McHardy G, Shull H, Kirsner JB. A co-operative double-blind evaluation of gastric 'freezing' in the treatment of duodenal ulcer. *N Engl J Med* 1969; **281**: 16–19. 10.1056/NEJM196907032810104
WorldCat Crossref

41. Bagenal FS, Easton DF, Harris E, Chilvers CE, McElwain TJ. Survival of patients with breast cancer attending Bristol Cancer Help Centre. *Lancet* 1990; **336**: 606–610. 10.1016/0140-6736(90)93402-B
WorldCat Crossref

42. Smithells RW, Sheppard S, Wild J, Schorah CJ. Prevention of neural tube defect recurrences in Yorkshire: final report (Letter). *Lancet* 1989; **ii**(8661): 498–499. 10.1016/S0140-6736(89)92103-X
WorldCat Crossref

43. Selikoff IJ, Churg J, Hammond EC. Asbestos exposure, smoking and neoplasia. *JAMA* 1968; **204**: 106–

112. 10.1001/jama.204.2.106
WorldCat Crossref

p. 50 44. Lindhout D, Schmidt D. *In-utero* exposure to valproate and neural tube defects. *Lancet* 1986; **i**: 1392–1393. 10.1016/S0140-6736(86)91711-3
Crossref

45. Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet* 2003; **362**: 419–427. 10.1016/S0140-6736(03)14065-2
WorldCat Crossref

46. Bingham S, Riboli E. Diet and cancer—the European Prospective Investigation into Cancer and Nutrition. *Natl Rev Cancer* 2004; **4**: 206–215. 10.1038/nrc1298
WorldCat Crossref

47. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. *BMJ* 1964; **i**: 1399–1410. 10.1136/bmj.1.5395.1399
WorldCat Crossref

48. Doll R, Gray R, Hafner B, Peto R. Mortality in relation to smoking: 22 years' observations on female British doctors. *BMJ* 1980; **i**: 967–971. 10.1136/bmj.280.6219.967
WorldCat Crossref

49. Hennekens CH, Speizer FE, Rosner B, Bain CJ, Belanger C, Peto R. Use of permanent hair dyes and cancer among registered nurses. *Lancet* 1979; **i**: 1390–1393. 10.1016/S0140-6736(79)92021-X
WorldCat Crossref

50. Stampfer MJ, Colditz GA, Willett WC, *et al*. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the Nurses' Health Study. *N Engl J Med* 1991; **325**: 756–762. 10.1056/NEJM199109123251102
WorldCat Crossref

51. Dawber TR. *The Framingham Study: The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press, 1980.
Google Scholar Google Preview WorldCat COPAC

52. Wilson PW. Established risk factors and coronary artery disease: the Framingham Study. *Am J Hypertens* 1994; **7**: 7S–12S.
WorldCat

53. Kuh D, Hardy R, Langenberg C, Richards M, Wadsworth ME. Mortality in adults aged 26–54 years related to socioeconomic conditions in childhood and adulthood: post war birth cohort study. *BMJ* 2002; **325**: 1076–1080. 10.1136/bmj.325.7372.1076
WorldCat Crossref

54. Wadsworth MEJ. *The Imprint of Time: Childhood, History, and Adult Life*. Oxford: Oxford University Press, 1991.
Google Scholar Google Preview WorldCat COPAC

55. Susser E, Terry MB. A conception-to-death cohort. *Lancet* 2003; **361**: 797–798. 10.1016/S0140-6736(03)12721-3
WorldCat Crossref

56. Barbour V. UK Biobank: a project in search of a protocol? *Lancet* 2003; **361**: 1734–1738. 10.1016/S0140-6736(03)13377-6
WorldCat Crossref

57. 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

58. Doll R, Hill AB. Smoking and carcinoma of the lung. Preliminary report. *BMJ* 1950; **ii**: 739–748. 10.1136/bmj.2.4682.739
WorldCat Crossref

59. Doll R, Hill AB. A study of the aetiology of carcinoma of the lung. *BMJ* 1952; **ii**: 1271–1286. 10.1136/bmj.2.4797.1271
WorldCat Crossref

60. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. *BMJ* 1976; **ii**: 1525–1536. 10.1136/bmj.2.6051.1525
WorldCat Crossref

61. MacGregor JE, Moss SM, Parkin DM, Day NE. A case–control study of cervical cancer screening in north east Scotland. *BMJ* 1985; **290**: 1543–1546. 10.1136/bmj.290.6481.1543
WorldCat Crossref

62. Selby JV, Friedman GD, Quesenberry CP, Jr, Weiss NS. Effect of fecal occult blood testing on mortality from colorectal cancer: a case–control study. *Ann Intern Med* 1993; **118**: 1–6.
WorldCat

63. Weiss NS, McKnight B, Stevens NG. Approaches to the analysis of case–control studies of the efficacy of screening for cancer. *Am J Epidemiol* 1992; **135**: 817–823.
WorldCat

64. Berwick M, Begg CB, Fine JA, Roush GC, Barnhill RL. Screening for cutaneous melanoma by skin self-examination. *J Natl Cancer Inst* 1996; **88**: 17–23. 10.1093/jnci/88.1.17
WorldCat Crossref

p. 51 65. Fentiman IS, Cuzick J, Millis RR, Hayward JL. Which patients are cured of breast cancer? *BMJ* 1984; **289**: 1108–1111. 10.1136/bmj.289.6452.1108
WorldCat Crossref

66. Taffler RJ. Forecasting company failure in the UK using discriminant analysis of financial ratio data. *J R Statist Soc* 1982; **145**: 342–358.
WorldCat

67. Pan SY, Johnson KC, Ugnat AM, Wen SW, Mao Y. Association of obesity and cancer risk in Canada. *Am J Epidemiol* 2004; **159**: 259–268. 10.1093/aje/kwh041
WorldCat Crossref

68. Redelmeier DA, Tibshirani RJ. Association betweeen cellular-telephone calls and motor vehicle collisions. *N Engl J Med* 1997; **336**: 453–458. 10.1056/NEJM199702133360701
WorldCat Crossref

69. Pauletto P, Puato M, Caroli MG, *et al*. Blood pressure and atherogenic lipoprotein profiles of fish-diet and vegetarian villagers in Tanzania: the Lugalawa study. *Lancet* 1996; **348**: 784–788. 10.1016/S0140-6736(96)01391-8
WorldCat Crossref

70. Törnberg SA, Carstensen JM. Relationship between Quetelet's index and cancer of breast and female genital tract in 47 000 women followed for 25 years. *Br J Cancer* 1994; **69**: 358–361. 10.1038/bjc.1994.65
WorldCat Crossref

71. Cartwright A. *Health Surveys in Practice and in Potential: A Critical Review of their Scope and Methods*. London: King Edward's Hospital Fund for London, 1983.
Google Scholar Google Preview WorldCat COPAC

72. Abramson JH. *Survey Methods in Community Medicine: Epidemiological Research, Programme Evaluation, Clinical Trials*,

5th edn. Edinburgh: Churchill Livingstone, 1999.
Google Scholar Google Preview WorldCat COPAC

73. Aday LA, Cornelius LJ. *Designing and Conducting Health Surveys: A Comprehensive Guide*, 3rd edn. San Francisco, CA:
p. 52 Jossey-Bass, 2006. ↳
Google Scholar Google Preview WorldCat COPAC