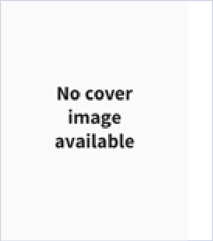


No cover image available

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

Search in this book

CHAPTER

5 Error and bias in observations

J. Mark Elwood

https://doi.org/10.1093/acprof:oso/9780198529552.003.05 Pages 123–156

Published: February 2007

Abstract

This chapter is divided into two parts. First, it discusses the general principles of identifying and minimizing error and bias in observations. Secondly, it looks at how to measure and adjust for bias. Self-test questions are provided at the end of the chapter.

Keywords: study design, source of error, observation bias, observation

Subject: Public Health, Epidemiology

Collection: Oxford Scholarship Online

> *Mathematics may be compared to a mill of exquisite workmanship, which grinds you stuff of any degree of fineness; but, nevertheless, what you get out depends on what you put in; and as the grandest mill in the world will not extract wheat-flour from peascod, so pages of formulae will not get a definite result out of loose data*
>
> —T. H. Huxley: Geological reform; 1869

This chapter falls into two parts. First we will discuss the general principles of identifying and minimizing error and bias in observations, and in the second part we will look at how to measure and adjust for bias.

Part 1. Identifying and minimizing error and bias: Sources of error and of bias

In the previous chapter we saw that the choice of the subjects for inclusion in the study defines one of the two key factors 'exposure' or 'outcome'. In cohort studies and intervention trials, the subjects are defined by their exposure or intervention, and the remaining factor to be assessed is the outcome. In the case–control approach, the subjects are selected by their outcome and the remaining factor to be assessed is the exposure. As we have seen, the way in which the subjects are chosen defines the study and determines its external validity. For example, a study of the value of physiotherapy in rheumatoid arthritis may be seen on closer examination to be relevant only to patients of a certain age who have a particular form of rheumatoid arthritis. The eligibility criteria and the participation rate will affect the external validity, i.e. the applicability and usefulness of the results in a wider context.

In the next stage of assessing scientific work, either our own or that of others, we accept what has been done in terms of the subjects included in the study and the design used. We can then ask this central question: Do the results support a causal relationship between the exposure and the outcome, within the confines of the particular study?

If any association is shown within the study, it must be due to one (or more) of four mechanisms: *observation bias*, *confounding*, *chance*, or *causation*. In this ↳ chapter we shall deal with observation bias. Observation bias is relevant to the measurement of the dependent variable in the study, i.e. the outcome in studies of a cohort design and the exposure in studies of a case–control design. p. 124

The central issue is the relationship between the *true value* of the factor being assessed, outcome or exposure, and the value of the variable that is chosen to represent that factor in the study. In the intervention study of physiotherapy and rheumatoid arthritis, the outcome might be defined as an improvement in the function of the affected joints. How this improvement can be best assessed will be a major component of the study design; possibilities range from physiological measures such as hand grip to questionnaire assessments of degree of functional impairment. Expert knowledge is obviously required, and attention must be paid to the acceptability, reproducibility, and relevance of the measures considered. The variable measured in a study is often considerably far removed from the biological factor or event that is defined in the causal hypothesis.

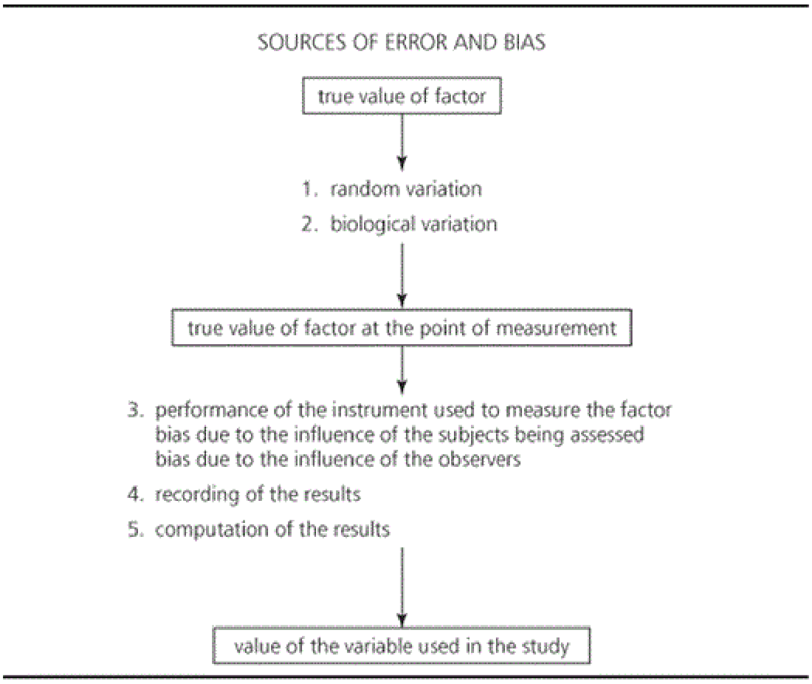
Consider a case–control study assessing whether high vitamin C consumption is protective against heart disease. The causal hypothesis relates the occurrence of heart disease to the intake of vitamin C over a long time period many years before the clinical diagnosis. The variable used to represent this factor in the retrospective design may be the responses to a questionnaire on the frequency of consumption of a number of food items at a defined period in the recent past, converted through a formula into an estimate of vitamin C consumption. The variable appearing in the results as 'exposure' is considerably different from the biological 'exposure' in the hypothesis.

A good example is given by studies assessing the association between gastric cancer and infection with *Helicobacter pylori*, a bacterium that can survive within the stomach. The causal hypothesis relates to the occurrence of stomach cancer being increased in subjects who have had *H.pylori* infection, perhaps over a long time period many years before the clinical diagnosis of this tumour. In a case–control study, with recently diagnosed cases of stomach cancer and controls, the easiest way to assess *H.pylori* infection is to use a laboratory test for the presence of the bacterium at the time of the study. Many such studies have been done, usually showing an association with a modest relative risk of around 2–4. However, past exposure to *H.pylori* may not persist, and so such tests give only an inaccurate estimate of past exposure. Patients with certain characteristics shown by immunological tests are much more likely to clear the bacteria from their stomach mucosa and so produce a negative test. This misclassification of exposure can be minimized by studying only

the subgroup of cases with immunological features likely to lead to persistence of the bacteria; one study using
p. 125 this method of control for misclassification ↳ showed increases in the observed odds ratio from 3.7, based on all cases, to 18.3, based on cases in which false-negative tests were unlikely [1].

Error: non-differential classification

There are several influences that may cause differences between the true value of the factor being assessed and the *recorded value* of the variable chosen; these are shown in **Ex.** 5.1. We need to distinguish two different problems. First is the problem of 'error', which means *inaccuracy that is the same in the different groups of subjects being compared*. This is also referred to as 'non-differential' misclassification—it does not differ between the different groups of subjects.


SOURCES OF ERROR AND BIAS
true value of factor
1. random variation
2. biological variation
true value of factor at the point of measurement
3. performance of the instrument used to measure the factor
bias due to the influence of the subjects being assessed
bias due to the influence of the observers
4. recording of the results
5. computation of the results
value of the variable used in the study

Ex. 5.1. Sources of error and of bias in the observed value of a variable compared with the true value of the factor it represents

Error includes several components (Ex. 5.1). If we assume that the true value is the usual value over a relevant time in an individual subject, the measurement used will have within-subject variation including true random variation, and also biological variations such as circadian or seasonal variations. All methods of measurement will have a degree of error as a function of the instrument used. Data have to be recorded, manually or
p. 126 electronically, for all ↳ studies, and in most studies raw data are converted in some way to give the final variable representing the factor under consideration.

To minimize error, methods of measurement and recording need to combine precision, reliability, and practicality, and the conduct of the study needs careful planning and monitoring. The details are specific to each topic and may be very complex; accurate assessments of exposures such as diet, electromagnetic fields, or social deprivation, and of outcomes such as cardiac function, mental state, or improved health, are all major topics in themselves. In general error is less of a problem than bias; therefore the general principle is that the *methods used will be applied in the same manner and with the same care to all the subjects in the study irrespective of the group to which they belong*. If this is done, then we will accept that a degree of error exists, but may be able to conclude that there is little possibility of systematic differences between the groups being compared.

Of course, error is important. The greater the error, the more ‘noise’ there is in the system, and therefore the more difficult it is to detect a true difference between the groups being compared. In the extreme situation, if the measurement used is so inaccurate that its value bears no relationship to the true value of the factor being assessed, we could not detect any differences between groups of subjects even if large differences exist. Thus, if physiotherapy is actually beneficial in improving joint function, a reasonably accurate method of assessing joint function will show this improvement, while a very inaccurate method will show no difference between treated and untreated groups of patients.

The effect of non-differential error will usually be to make the observed association closer to the null value than is the true situation. Except in some unusual situations, if a study shows a strong association, this association cannot be produced by error in the measurements used; on the other hand, if a study shows no association or a weak association, error in the observations may be disguising a much stronger association.

Studies of disease causation often produce only weak associations, with relative risks of the order of 2. The reason may be that the exposure variable measured is a very inaccurate estimate of the true biological factor concerned, because it is an indirect measurement. Many studies show that the incidence of breast cancer is increased in association with obesity, with fairly low relative risks, and this may be because obesity is an indicator, but an inaccurate indicator, of a specific dietary factor which is related to breast cancer incidence. Where the factor assessed is a closer estimate of the true biological agent, relative risks will be higher. The inhalation of certain types of wood dust is a cause of cancers in the nose and nasal sinuses; if we compare p. 127 employees in an industry ↳ which uses wood with the general working population, we find a moderately increased relative risk, perhaps of 2–3; if we compare workers employed on dusty processes which use wood with the general working population, we find a much higher relative risk, perhaps 10 or more, while if we assess workers who personally have had exposure over many years to particular types of wood we find a relative risk of 100 or more. Thus in a study in France, a relative risk of 303 was found for workers with over 35 years exposure to hardwoods, while exposure to softwoods showed no increased risk [2]. The closer we come to the biological causal factor, the higher the relative risk will become.

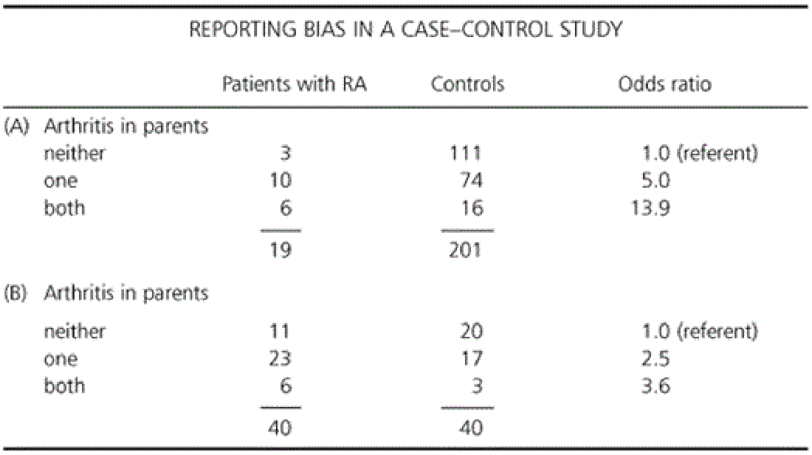
The quality of information may depend on its source. A study in Vermont, USA, compared the data obtained from interviews with 857 men and their wives with regard to the wife’s pregnancy history. There was disagreement on the numbers of live born children or on their dates of birth in 11 per cent of couples. Taking the history given by the wife as being accurate, husbands reported only 70–74 per cent of spontaneous abortions and low-birth-weight births, and only 35 per cent of induced abortions [3].

Bias: differential misclassification

The other component of inaccuracy is bias, i.e. inaccuracy that is different in its size or direction in one of the groups under study than in the others. This is a much more serious problem, as bias can influence the results of a study in any direction. It can produce measurements of association that are exaggerated, and may produce strong associations when there is no true difference between the groups being compared.

Examples of bias

Some quite dramatic examples of observation bias can be found. In a study to assess whether rheumatoid arthritis has a familial link, patients with the disease (cases) were asked whether their parents had suffered from arthritis, and their responses were compared with those of unaffected controls [4]. The results showed that the frequency with which parents were affected was much higher for the rheumatoid arthritis patients than for controls, with a high relative risk (**Ex.** 5.2). In a second comparison, another group of patients with rheumatoid arthritis were asked about arthritis in their parents, and independently the unaffected siblings of these rheumatoid arthritis patients were asked the same questions. Of course, the answers relate to the same parents, and therefore should be identical; however, these results also show that a higher frequency of parental p. 128 arthritis was reported by the rheumatoid arthritis ↳ patients than by their unaffected siblings, and this result must be due to observation bias. Patients affected by a disease are more likely to know of family members with the same disease.

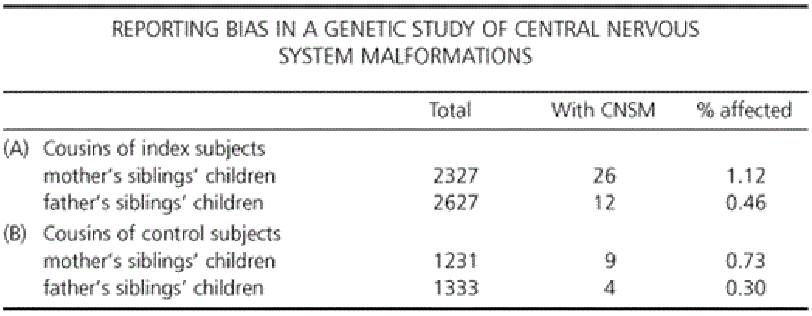


REPORTING BIAS IN A CASE–CONTROL STUDY

	Patients with RA	Controls	Odds ratio
(A) Arthritis in parents			
neither	3	111	1.0 (referent)
one	10	74	5.0
both	6	16	13.9
	19	201	
(B) Arthritis in parents			
neither	11	20	1.0 (referent)
one	23	17	2.5
both	6	3	3.6
	40	40	

Ex. 5.2. An example of subject recall bias. Results of case–control studies comparing patients with rheumatoid arthritis (RA) with unaffected controls in terms of whether their parents had arthritis, as reported by these respondents. Study A shows a strong positive association between RA and arthritis in parents. Study B also shows a positive association which, as it compares responses of RA patients and their unaffected siblings with regard to the same parents, must be due to variation in reporting. Knowledge of the results of study B will influence the interpretation of study A. From Schull and Cobb [4]

Studies of family history which depend upon asking mothers of babies with severe defects whether their relatives have had similarly affected babies have sometimes resulted in the conclusion that the disease occurred more commonly in the maternal than the paternal relatives. Such observations have led to hypotheses of complex inheritance patterns. These observations may be biased by the fact that mothers generally know more about the offspring of their own family than that of their husband's family, as can be seen if control families are investigated in the same way (**Ex.** 5.3). These data from a large family history study show that in the cousins of subjects with central nervous system malformations the reported frequency of similarly affected children was considerably higher in the mothers' relatives than in the fathers' relatives, suggesting a maternal inheritance pattern. However, the same type of investigation carried out on families of normal (control) babies showed a very similar degree of excess of malformations in the mothers' families compared p. 129 with the fathers' families, which must be due to biased reporting [5]. These examples emphasize ↳ that if the information is biased, the associations seen may be strong, and there is no point in applying statistical tests to biased data. The fact that the associations are statistically significant gives no protection against bias.



REPORTING BIAS IN A GENETIC STUDY OF CENTRAL NERVOUS SYSTEM MALFORMATIONS

	Total	With CNSM	% affected
(A) Cousins of index subjects			
mother's siblings' children	2327	26	1.12
father's siblings' children	2627	12	0.46
(B) Cousins of control subjects			
mother's siblings' children	1231	9	0.73
father's siblings' children	1333	4	0.30

Ex. 5.3. Recall bias in a genetic study. Table A compares the reported frequency of central nervous system malformations (CNSM) in cousins of an index series of 547 cases of these defects, and shows higher frequencies of CNS defects in maternal compared with paternal relatives (relative risk 2.4). However, Table B shows the reported frequencies of CNSM in cousins of control births which did not have CNS defects, and shows a similar maternal–paternal difference (relative risk 2.4). From Carter *et al*. [5]

A notorious issue of observation bias arose with regard to a randomized trial comparing antibiotics and placebo in the treatment of otitis media in children, conducted in Pittsburgh, which resulted in a major conflict referred to as the 'Cantekin affair'. The first report of this trial, published in the *New England Journal of Medicine*, showed a doubling of the frequency of clinical resolution of the disease with antibiotic treatment, based on an outcome determined mainly by clinical examination [6]. However, other investigators in the trial thought that the ear examinations were open to observation bias, and submitted another analysis concluding that no benefit was seen, based on more objective tympanometric measurements [7]. They concluded that the clinical examination results for each ear were not done independently, and that they were not compatible with more objective measurements. In fact, the difference was more quantitative than qualitative: the clinical examination showed a larger effect, while the results based on tympanometry also showed a benefit, but smaller, and not statistically significant in the analysis published. Both groups of investigators agreed that no benefit was shown if a further endpoint, hearing tests, was used. There were allegations that some of the clinical investigators had been influenced by drug company funding. This conflict led to several investigations, which had criticisms of both main parties, and raised issues about the role of journal editors [8,9].

p. 130

Methods to minimize bias

The most important sources of bias are variation in the subject's response to the method of assessment, and variation in the observer's response (Ex. 5.1). The main principle in avoiding bias is to ensure that the same methods are used under the same circumstances by the same observers for all subjects involved in the study, and to employ double- or single-blind techniques as far as possible. In the study design the choice of outcome or exposure measures is important, and these must not only be relevant to the hypothesis, but be chosen to be objective, reproducible, and robust, i.e. likely to be little influenced by variations in the method of testing. We must guard against mistakes in both directions, however; while an outcome which is extremely difficult to measure and is open to highly subjective interpretation may be of little value, there is also the danger of choosing an outcome simply because it can be measured easily, even if it is not directly relevant to the hypothesis under test, or may even result in a distortion or change in that hypothesis. For example, we may want to know if a health education programme results in subjects changing their diet; but as this is very difficult to measure, we may choose to use something much simpler, such as the subjects' responses to factual questions about diet. This is perfectly appropriate if we accept that the hypothesis under test has now changed, and we are assessing only whether the educational programme results in increases in knowledge. However, we must not assume, unless we have good evidence, that an improvement in knowledge will be linked to a change in behaviour.

Single- and double-blind methods

The best method of avoiding bias in measurements is to ensure that neither the subject nor the observer is aware of which group the subject is in, so that the study is *double-blind*. This is easiest to do in a prospective intervention trial. Thus to evaluate the effect of a new drug on the relief of symptoms from rheumatoid arthritis, a placebo drug can be made up which looks and tastes the same as the new drug, and the study is designed so that neither the subject nor the person making the observations of outcome is aware of whether the subject is taking the active drug or the placebo. Such a design should avoid most sources of subject and observer bias. Care has to be taken that the double-blindness is in fact preserved, and not broken either inadvertently through administrative lapses or through the active drug having some feature or side effect that makes its presence obvious. In the trial of treatment for acne described in Chapter 4 [10], the laser treatment procedure involved moving the laser instrument over the face, with the patients wearing opaque eye
p. 131 protection. The same procedure was followed for the controls, but with ↳ the instrument disconnected, to achieve patient blindness. The treatment outcome was assessed by a standardized examination with counting and classification of lesions, with the assessors unaware of which treatment had been given, to achieve observer blindness. The treatment was not usually perceptible, but two patients reported discomfort and would have been aware that they were receiving active treatment, so blindness cannot be assured.

In many other prospective designs, double-blindness is not possible. For example, if a comparison is being made between surgery and medical therapy for coronary artery disease, it is impossible to achieve double-blindness. A *single-blind* assessment, where the observer is unaware of the treatment given to the subject, may be possible. While the normal medical carers of the patient will be aware of the patient's treatment, outcome measures could be chosen as items which can be verified by an independent group of assessors on the basis of electrocardiograms, radiographs, and so on, or those in which observer bias is not an issue, e.g. total mortality. In the 1948 tuberculosis trial, those assessing the radiographs were kept unaware of the treatment the patient had received [11].

A randomized trial comparing specialist- and GP-led prenatal care was described in Chapter 4 [12]. There was potential observation bias in this trial. The trial obviously could not be single- or double-blind. Most of the clinical outcome information was based on routine clinical records, and so the validity of recording may differ in the two groups. For example, fewer women in the GP care group were recorded as having hypertension, proteinuria, or pre-eclampsia. This could be the real situation, despite the randomization, but could also indicate relative under-diagnosis or under-recording. The two groups had similar numbers of women with previously undiagnosed hypertension recorded at the time of admission in labour, giving some protection against this bias. An important section of the results is the self-report on satisfaction. This could be biased, as of course the women are only aware of the care system they have experienced. An important aspect is that the response rate to this questionnaire was quite good (78 per cent), and was virtually identical in each of the two groups.

The term 'triple-blind' has been used where the analysis of the trial is carried out by investigators with information on which study group each subject is in, without knowing which treatment was allocated to each group, which is often easy to achieve and is desirable. It is particularly important in interim analyses, which may modify the conduct of the study and even terminate it early. An example is a trial of homeopathic treatment for chronic fatigue syndrome, where control of potential bias is obviously important [13]. To add to
p. 132 the ↳ confusion, some authors separate blindness of the investigators who deal with entry to the study and treatment from those investigators who determine outcome; then, adding patient blindness and data analyst blindness produces a 'quadruple-blind' study; an example is a trial of prevention of renal dysfunction after cardiac surgery [14]. Less seriously, some have referred to higher levels of blinding as studies where, even after analysis, no-one knows what the results mean.

Bias in cohort studies

In cohort studies, the subjects are selected in terms of their exposure, and the bias question applies to the outcome data. In observational cohort studies the subjects are usually aware of their exposure and the outcome may be assessed by the subjects' response to questionnaires or by routine clinical records. In such a situation, single- or double-blind outcome assessment may be impossible. It may be useful to compare the exposed and comparison groups in terms of the frequency with which routine examinations are done or outcomes are reported, who reports them, and the completeness and consistency of the observations. Comparability in these process measures will support comparability in the results. Another useful ploy is to look for specificity of the result, by showing that outcomes that are irrelevant to the causal hypothesis are similar in the groups being compared.

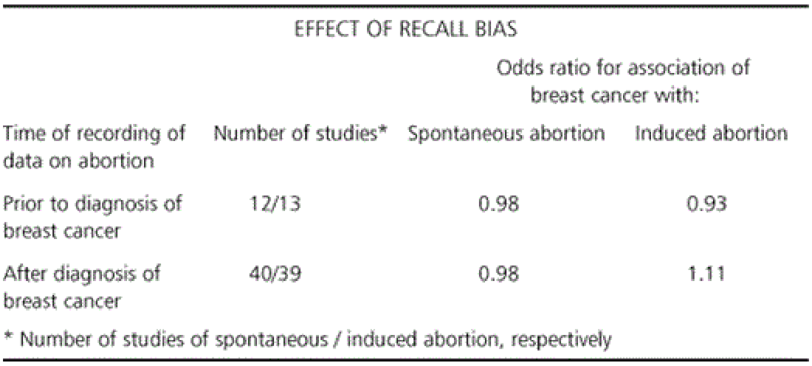
Two cohort studies of oral contraceptive use were described in Chapter 4 [15,16]. Both these studies were weak in terms of potential observation bias, as the outcome data were based on routine medical and clinical records. These could have been biased directly by knowledge of the method of oral contraceptive use, if certain conditions were looked for more carefully in women using a certain type of contraceptive method, and also indirectly, in that women using oral contraceptives might have visited a general practitioner or clinic more, or less, frequently than other women. For example, in the Royal College of General Practitioners' study, 18 per cent of all diagnoses were recorded on the prescription date of the oral contraceptive, suggesting that some complaints might have come to the general practitioner's notice only because the patient had to visit for the prescription [15]. Observation bias will be less likely if the assessment methods are more objective. In the Family Planning Association study, the association between oral contraceptive use and venous embolism was stronger where the evidence for the diagnosis was more objective [17], making bias less likely as an explanation. In this study, the morbidity information used was restricted to hospital referrals for inpatient or outpatient assessment, or inpatient or outpatient care, partly in order to avoid problems of observation bias.

Bias in case–control studies; recall bias

As in a cohort study, in case–control studies the accuracy of case definition may need to be considered. For example, in a case–control study of pulmonary embolism and deep venous thrombosis, cases were categorized as having definite, probable, or possible venous thromboembolism on the basis of the diagnostic information available. The observed increased risk in association with the use of hormone replacement therapy was higher for definite and probable diagnoses than for possible diagnoses, which is consistent with a true effect [18].

In case–control studies, subjects are selected in terms of the outcome, and the main bias issue applies to the documentation of past exposure. In most case–control studies, information is obtained by interviewing cases and controls. The central issue is *recall bias* (or response bias), a differential response to questions between cases who have been diagnosed with disease and controls who have not. A good study design will ensure use of a well-designed standardized interview, a consistent approach by well-trained interviewers, and a supportive and non- judgemental atmosphere for the interview. However, a difference in the ability or willingness to report past events is likely, even if unconscious, on the part of the subject. Where the study concerns sensitive issues, this recall bias may be more marked.

For example, a meta-analysis (by methods to be described in Chapter 8) has brought together data on 83 000 women with breast cancer from 53 studies in 16 countries, relating breast cancer to a previous spontaneous or induced abortion (**Ex.** 5.4). These included cohort studies and case–control studies in which record linkage methods used information on abortion that had been recorded before the occurrence of the breast cancer; these studies may have random error in the data on abortions, but differential bias can be excluded. There were also case–control studies using retrospective interviews, where the data on abortions were collected from the women after the diagnosis of breast cancer in the cases; these studies are open to bias as well as random error. For spontaneous abortion, both types of study showed no association. For induced abortion, no increased risk was seen in cohort studies or in the case–control studies with information on abortion recorded before the occurrence of the breast cancer. However, the case–control studies using retrospective interviews showed a modest but statistically significant association with induced abortion which, given that no association was seen in the other studies, is due to recall bias. The women who have been diagnosed with breast cancer must have reported induced abortions more readily than the control women [19].



EFFECT OF RECALL BIAS

		Odds ratio for association of breast cancer with:	
Time of recording of data on abortion	Number of studies*	Spontaneous abortion	Induced abortion
Prior to diagnosis of breast cancer	12/13	0.98	0.93
After diagnosis of breast cancer	40/39	0.98	1.11

* Number of studies of spontaneous / induced abortion, respectively

Ex. 5.4. Effect of recall bias. Results from a combined analysis of studies of breast cancer and both spontaneous and induced abortion. From Collaborative Group on Hormonal Factors in Breast Cancer [19]

p. 134 A degree of blinding may be accomplished by doing the study on subjects with symptoms considered suggestive of the outcome under test, but who have not received a specific diagnosis. In the early study of lung cancer and smoking mentioned in Chapter 2 [20], bias in the smoking histories is made less likely because the smoking histories of patients thought to have lung cancer at the time of interview, but in whom the final diagnosis was not lung cancer, were similar to the controls and differed from the histories in confirmed cases. To study psychological factors in women with breast cancer, carrying out the interviews in screening clinics before examination allows comparison of women with breast cancer with those with normal results without the biases that might result from knowledge of the diagnosis. A control group may be chosen from subjects who have conditions of similar severity or nature to that of the cases; thus in a retrospective study looking at events during pregnancy in mothers of infants born with a specific congenital defect, the control group could be chosen as mothers of babies with a range of other defects, who would be expected to be influenced by the same factors affecting recall.

Observer variation is also difficult to control in a case–control study, as it is not easy to keep the observer unaware of the status of the interviewee. Simple precautions such as ensuring that cases and controls are assessed by the same observer or group of observers, and by the same methods used under the same circumstances, are of help. Some observations may be helpful in judging whether subject or observer bias may be a problem, such as recording the length of time taken for examinations or interviews, recording the interviewer's assessment of the cooperation of the subject and the degree of difficulty experienced with some of the key questions, and asking the subjects at the end
p. 135 ↳ of the interview whether they are aware of any relationship between their condition and some of the factors asked about. Similarly, the examiners or interviewers can be asked to record whether they became aware of the case or control status of the subject before or during the assessment. Such recordings give the possibility of analysing subsets of data for subjects who were aware or were not aware of the key hypothesis, and those in whom the observer did or did not know their status. Questions for which cases and controls would be expected to give similar answers may be useful.

If the study relies on a small number of interviewers, the results for each interviewer should be examined. In a large survey of women in the USA carried out by four different interviewers, no interviewer variation was seen for questions requiring recall of specific events, but the responses to questions involving subjective and personal information or requiring further probing from the interviewer varied between interviewers. As a result, results on the impact of support networks on psychological symptoms varied depending on which interviewers' data were used [21].

In the case–control study of breast cancer in New Zealand described in Chapter 4 [22], the information was collected by a standardized telephone interview, after an initial approach by letter. The standardization of the interview, and the fact that the interviewer did not know whether the interviewee was a cancer patient or a comparison subject, provides some protection against bias on the part of the interviewer. However, the major likely source of systematic observation bias is from the subjects themselves, who of course were well aware of whether they had been treated for breast cancer or not. A standardized, non-emotive, and systematic interview technique is the best protection against such bias. The bias could also be overcome if information on the exposure of interest, in this case oral contraceptive use, were obtained from other independent sources, such as medical records. However, doing this is often difficult in practice, and it is also unlikely that such sources would give comparable information on all the relevant confounding factors. In this study the investigators did assess general practitioners' records for women who reported recent use of prescribed contraceptives, and concluded that there was 'close agreement' between this information and that given by the women themselves.

Practical issues in reducing bias and error

The design of methods of investigation that minimize error and bias is a large subject in its own right, and will not be dealt with fully here. However, it is useful to summarize some of the main approaches. Important issues include the *definition* of the items to be recorded, the choice of *methods of measurement*, the *standardization* of procedures, and *quality control* of all aspects of data ↳ gathering and processing. It is essential in any research study to define precisely the factor being assessed, even, one might say especially, when it appears simple; consider the definitional issues involved in items such as tumour stage, cardiac failure, pain relief, social class, diastolic blood pressure, high fat diet, or cellular atypia. p. 136

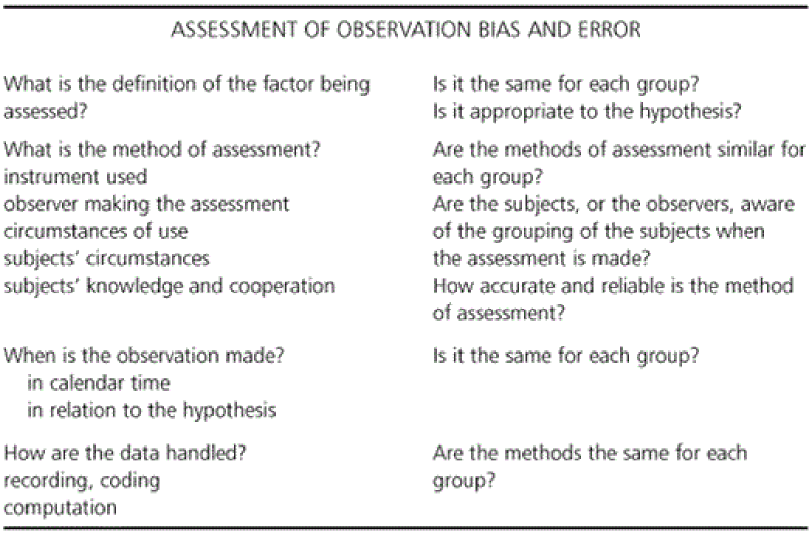
The 'instrument' used to assess the factor must then be chosen; we use this general term to include any means of assessment, such as a clinical examination, laboratory test, questionnaire, review of medical records, or observation. The way in which the instrument is to be applied must be standardized: by whom, when, how, and under what circumstances.

As an example of a difficult item to measure, the prevalence of stress disorder in 641 Australian Vietnam veterans was assessed. The lifetime prevalence of combat-related post-traumatic stress disorder assessed by a standardized interview format was 12 per cent, but when assessed by an interview which gave the interviewer the opportunity to interact more with the subject it was 21 per cent. Moreover, with the latter instrument, the prevalence found when the interviewer was a trained counsellor was up to twice as high as with non-counsellor interviewers, and was considerably greater for a female counsellor than for a male [23].

Quality control procedures should be developed to monitor the information collected throughout the study, and to produce useful data that will attest to its quality. These processes of definition, standardization, and quality control are relevant not only to the collection of information, but also to recording, coding, and computer entry. Quality control should include systematic checks for gross errors such as variables which are irreconcilable (e.g. males with menstrual problems), contradictory (e.g. non-matching age and date of birth), or outside an expected range, and systematic checks for inconsistencies, such as addresses or diagnoses from different sources; and rechecking of all or a sample of examination, interview, coding, and data entry procedures.

Assessment of bias and error in a completed study

A checklist for the assessment of bias and error is given in **Ex** 5.5. The questions can assess whether the variable recorded is a valid measurement of the biological factor that is relevant to the hypothesis under test. For each aspect of this measurement, a general question relates to the accuracy of the mea surement, i.e. primarily to error, and a more specific question deals with differences in the assessment between the groups under test, and therefore relates primarily to bias. Bias is the more serious problem as it may affect internal validity; we concentrate on whether the methods used have been applied identically to all subjects in the study, and whether the subjects' responses to these methods are likely to have been similar. Error will always be
p. 137 present, but may ↳ be a serious problem, causing the results of the study to be influenced towards a null result. Often the assessments can only be made in a qualitative and subjective manner. We need to make a reasoned judgement as to whether the results can be accepted as valid. Good evidence on the validity, reproducibility, or consistency of the observations made is helpful, and so opportunities for such assessments should be taken when designing a study.



ASSESSMENT OF OBSERVATION BIAS AND ERROR

What is the definition of the factor being assessed?	Is it the same for each group? Is it appropriate to the hypothesis?
What is the method of assessment? instrument used observer making the assessment circumstances of use subjects' circumstances subjects' knowledge and cooperation	Are the methods of assessment similar for each group? Are the subjects, or the observers, aware of the grouping of the subjects when the assessment is made? How accurate and reliable is the method of assessment?
When is the observation made? in calendar time in relation to the hypothesis	Is it the same for each group?
How are the data handled? recording, coding computation	Are the methods the same for each group?

Ex. 5.5. Bias and error. An outline scheme to assist in the consideration of issues of observation bias and error. The questions should be considered for the whole study, and specifically with regard to the comparability of the relevant groups: exposed and unexposed in cohort and intervention studies; affected and unaffected in case–control studies

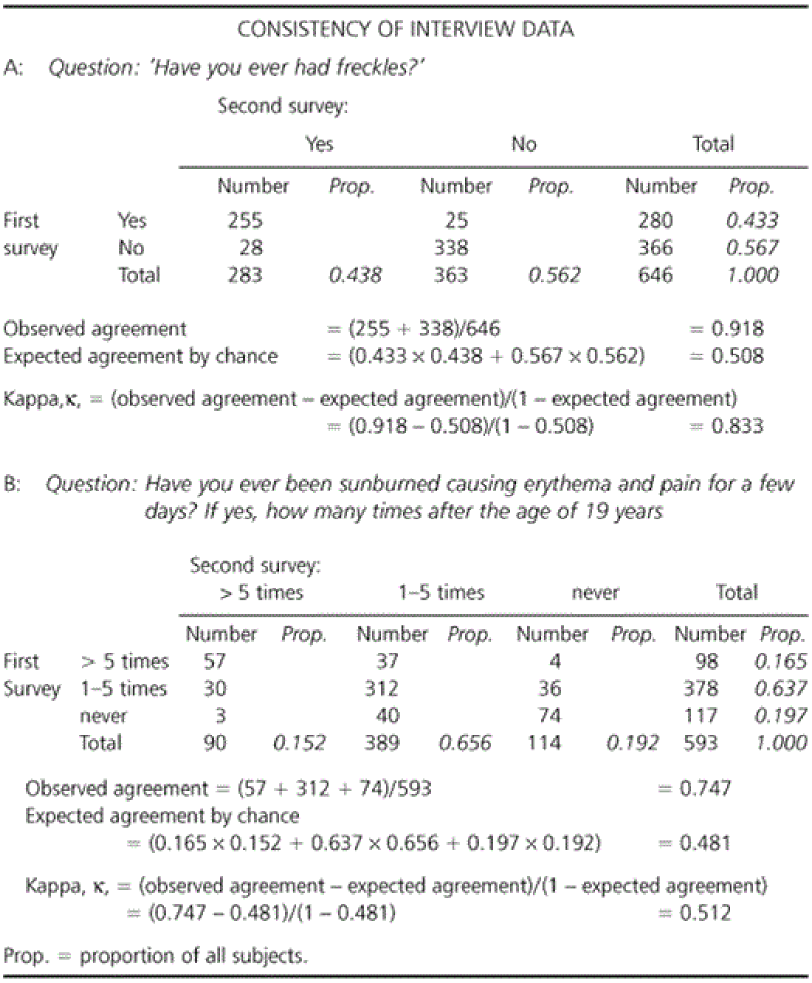
Part 2. Measuring aspects of observation bias, and controlling for bias: Measuring the consistency of information

It is important not only to collect data of good quality, but to be able to demonstrate its quality. The first criterion is *consistency*: data collected on more than one occasion or by more than one method or observer should be consistent. Checks for consistency can be made within an observation procedure such as an interview or clinical examination by repeating key items at different points in the procedure, and by assessing factors in more than one way. All or a sample of subjects can be reassessed; the likelihood of a true change over
p. 138 time must be considered in interpreting the results. Observer and instrument ↳ variation can be assessed on all or a sample of subjects, comparing data collected independently by a number of observers or by different methods. To capitalize on some of these techniques variables can be deliberately included in the study, chosen not because they are of intrinsic interest but because they should be stable over time, between interviewers, between subjects, and so on, and therefore variation in their recorded values can be used to assess the consistency of the information.

A quantitative measure of consistency: kappa

This quantitative method of assessing consistency is applicable where two methods have been used on the same subjects (two observers, two occasions, or two procedures). **Exhibit** 5.6 shows some data from a study that assessed the reproducibility of a self-administered questionnaire relating to risk factors for melanoma (a skin cancer) [24]. The data relate to subjects completing two questionnaires 1–3 years apart. For the question you ever had freckles?’ (part A), 593 of the 646 respondents replied in the same way for each questionnaire (255 as ‘yes’, 338 as ‘no’), giving an overall proportion with consistent responses *a* = 593/646 = 0.918. This quantity may be misleading, as even if there were no relationship between an individual’s responses to the two surveys, substantial agreement would be expected by chance alone; this amount can be calculated in the manner shown. The logic is that as 0.433 of all subjects gave a ‘yes’ response on the first survey, and 0.438 on the second survey, if the two responses were unrelated, the expected proportion giving ‘yes’ responses on both occasions would be 0.433 × 0.438 = 0.190. Similarly, the expected proportion responding ‘no’ on both surveys is 0.567 × 0.562 = 0.319. The total expected agreement is the sum of the expected agreement for each category; here it is 0.508. If we take the excess of agreement over expected agreement by chance (0.918 − 0.508), and divide it by the potential excess, which is (1 − 0.508), we obtain a statistic known as kappa (κ) [25]:

$$\kappa = (a - e)/(1 - e)$$



CONSISTENCY OF INTERVIEW DATA

A: *Question: 'Have you ever had freckles?'*

Second survey:

		Yes		No		Total	
		Number	*Prop.*	Number	*Prop.*	Number	*Prop.*
First	Yes	255		25		280	*0.433*
survey	No	28		338		366	*0.567*
	Total	283	*0.438*	363	*0.562*	646	*1.000*

Observed agreement = (255 + 338)/646 = 0.918

Expected agreement by chance = (0.433 × 0.438 + 0.567 × 0.562) = 0.508

Kappa, κ, = (observed agreement − expected agreement)/(1 − expected agreement)
= (0.918 − 0.508)/(1 − 0.508) = 0.833

B: *Question: Have you ever been sunburned causing erythema and pain for a few days? If yes, how many times after the age of 19 years*

Second survey:

		> 5 times		1–5 times		never		Total	
		Number	*Prop.*	Number	*Prop.*	Number	*Prop.*	Number	*Prop.*
First	> 5 times	57		37		4		98	*0.165*
Survey	1–5 times	30		312		36		378	*0.637*
	never	3		40		74		117	*0.197*
	Total	90	*0.152*	389	*0.656*	114	*0.192*	593	*1.000*

Observed agreement = (57 + 312 + 74)/593 = 0.747

Expected agreement by chance
= (0.165 × 0.152 + 0.637 × 0.656 + 0.197 × 0.192) = 0.481

Kappa, κ, = (observed agreement − expected agreement)/(1 − expected agreement)
= (0.747 − 0.481)/(1 − 0.481) = 0.512

Prop. = proportion of all subjects.

Ex. 5.6. Use of a re-interview technique to assess consistency of data. For questions related to the aetiology of melanoma, subjects were re-interviewed after a period of 1–3 years, and the data on consistency were used to adjust the risk estimates in a case–control study. From Westerdahl *et al*. [24]

where *a* is the proportion of subjects giving consistent responses and *e* is the proportion with consistent responses expected by chance alone. Kappa has a range of +1 (complete agreement), to 0 (agreement equal to that expected by chance), to negative values (agreement less than that expected by chance). Here kappa is 0.83, which indicates good consistency.

Another question was about sunburn, with the number of times it occurred categorized into three groups (part B). The observed agreement is the proportion of subjects who give the same response in both surveys and so
p. 139 appear on the diagonal of the table, and is 57 + 312 + 74 + 443/593 = 0.747. ↳ This compares with the chance agreement of 0.481, calculated as shown, yielding a kappa value of 0.51. Thus there is much less consistency in the response to this question than there was for the question on freckles. However, there are two complications with this question. First, subjects who responded 'never' on the first survey and 'more than five times' on the
p. 140 second survey, or vice versa, show a greater level of disagreement than subjects whose responses were in ↳ adjacent categories. The calculation shown is a simple unweighted calculation that does take into account different types of disagreement. A more appropriate formula is for *weighted kappa*, where 'weights' can be assigned to account for different degrees of disagreement, as described in texts such as Fleiss [26] and in many software packages. The weighted kappa result from this study was 0.54, somewhat higher than the value given by the simpler calculation. The other issue is that the true result may have changed; as the second survey was conducted later, more sunburns may have occurred, and we might expect a shift to the reporting of more sunburns in the second survey. However, examination of the overall proportions from the two surveys does not

show any shift in this direction; in fact, the proportion of subjects reporting large numbers of sunburns in the second survey is slightly lower than in the first survey, and so a real augmentation of experience is unlikely to explain the discrepancy in the interview results.

The variance of kappa can be calculated and used to test whether the observed agreement is statistically significantly better than that expected by chance alone. Confidence limits can be calculated (see Chapter 7 for a general explanation of confidence limits, and Appendix Table 10 for their calculation). The index can also be generalized to apply to studies with many categories of result and many observers, and to take account of different degrees of disagreement.

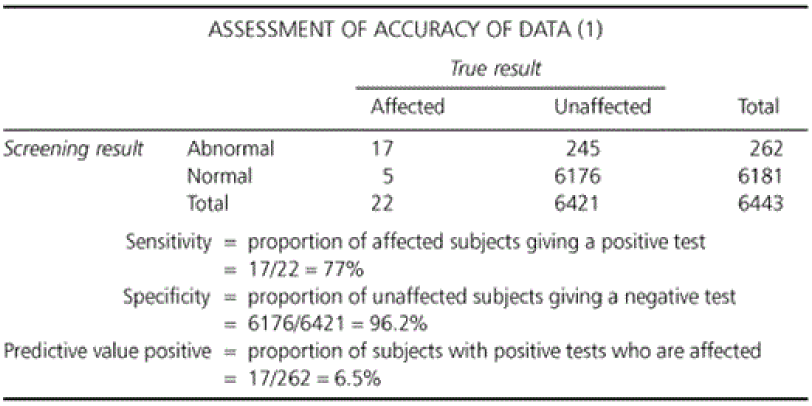
These measures of agreement have several limitations. First, their interpretation is subjective. Harlow and Linet [27] reviewed the agreement between questionnaire data and medical records for a large number of items, and suggested that kappa values above 0.80 should be regarded as showing very good agreement, 0.60–0.80 good agreement, 0.40–0.60 moderate, and values under 0.40 fair to poor agreement. This simple categorization has been used frequently but is only a rough guide; it is more important to assess the impact of the degree of inconsistency. The consistency of some common clinical assessments has been shown to be quite low, with kappa values of 0.3–0.7; many are in the range 0.4–0.6. Secondly, the kappa value will vary with the prevalence of the condition, being difficult to interpret where the prevalence is very low or very high [28,29]. If the prevalence is very low, kappa may be low even if the consistency is high; an example will be shown in the next section.

However, consistency results may indicate the best of several ways of assessment. In studies of the value of routine skin screening, the quality of information obtained by interview is important. In a study in Australia, subjects were asked if they had had an examination of their skin by a doctor in the last 3 years, and in another question if they had had an examination in the last 12 months, and the responses were checked against doctors' records. The kappa p. 141 ↳ value for the 3-year question was 0.87, showing very good agreement, but it was only 0.27 for the question on 12 months. This difference was attributed to *telescoping*, i.e. patients tend to remember events as being more recent than they were; many subjects reported an examination within the last 12 months when in fact it had been more than 12 months in the past. This comparison led to a decision that an analysis based on reported three-yearly screening would be better for further studies [30].

Assessment of the accuracy of information: sensitivity, specificity, and predictive value

If one method of assessment can be regarded as definitive, often termed a *gold standard*, the accuracy of any other method can be assessed against it. For example, a screening or diagnostic test can be compared against the final diagnosis achieved after full investigation. A measure of overall consistency is not so useful here, as the consequences of a positive and a negative result will be very different.

Sensitivity and *specificity* together describe the performance of a test against the 'true' result, but can be calculated only where the true result is known for all subjects. The *sensitivity* of the test measures its accuracy in identifying truly affected subjects. In the example shown in **Ex.** 5.7, the sensitivity is 77 per cent: the screening test identified 77 per cent of all affected subjects; the other 23 per cent gave a normal test result, and were *false negatives* [31]. *Specificity* is the accuracy in identifying subjects who are truly unaffected; here the specificity is 96.2 per cent, meaning that 96.2 per cent of unaffected subjects had a normal screening test result. The other 3.8 per cent were *false positives*.



ASSESSMENT OF ACCURACY OF DATA (1)

		True result		
		Affected	Unaffected	Total
Screening result	Abnormal	17	245	262
	Normal	5	6176	6181
	Total	22	6421	6443

Sensitivity = proportion of affected subjects giving a positive test
= 17/22 = 77%
Specificity = proportion of unaffected subjects giving a negative test
= 6176/6421 = 96.2%
Predictive value positive = proportion of subjects with positive tests who are affected
= 17/262 = 6.5%

Ex. 5.7. Assessment of accuracy of data by comparison with a fully accurate method: the validity of antenatal screening for neural tube defects by a measurement of -fetoprotein in maternal serum at 16–22 weeks' gestation compared with the presence of a neural tube defect assessed after delivery. The test is only the first step in the screening process; the ultimate result was that terminations were carried out on 16 affected and two unaffected pregnancies. For open neural tube defects, the sensitivity was 17/18 (94 per cent). From Wald *et al*. [31]

A kappa statistic is of little value here. It only assesses overall consistency, whereas the implications of sensitivity and specificity are quite different. Also, the simple kappa statistic is not useful where prevalence is low. In Ex. 5.7, 96 per cent of subjects overall are classified correctly, but the low prevalence means that the likelihood of correct classification by chance is also high (95.6 per cent). Calculation of the kappa value gives a result of 0.11, which is not very informative.

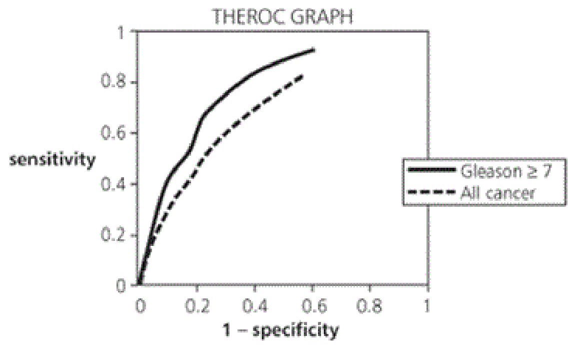
In routine screening and diagnostic applications, subjects giving a positive result to the first test will be investigated further; thus the *predictive value positive* will be the most easily measured parameter. It is a very relevant one, showing the proportion of all those testing positive, and suffering the consequences of worry and further testing, who do have the condition being sought. Here, 262 subjects had positive tests, which required further investigation, and 17 (6.5 per cent) were confirmed as positive. The predictive value increases with both
p. 142 the ↳ sensitivity and the specificity of the test, and, for given levels of these, also increases as the true prevalence of the condition in all those tested increases. The *predictive value negative*, the proportion of all those with a negative test who in fact are disease free, may need specific research to measure it, as all persons who test negative will need to be further assessed so that those with disease can be identified. In the example, the predictive value negative is 6421/6443 (99.66 per cent). However, note that, because of the low prevalence of disease, this impressive specificity still means that there are 245 mothers with false-positive results, compared with only 17 true positives; all 262 need further investigation.

A test that is very sensitive will miss few cases of disease; it has a high negative predictive value, i.e. a negative result is reliable. This has been described in clinical terms by the mnemonic SNOUT: if a Sensitive test is Negative it rules OUT disease. Equally, a test that is very specific means that few people without disease will have positive tests and so the positive predictive value is high, i.e. a positive test is reliable. The mnemonic is SPIN: if a SPecific test is Positive, it rules IN disease. The ideal balance depends on the purpose of the test. For example, to test donated blood for HIV virus, we need to know that a negative test result will give virtual
p. 143 certainty of the safety of the sample. ↳ The test has to be highly sensitive to rule out a hazard. However, if we were testing an individual patient for HIV, the most important thing is to avoid giving a false-positive report, and so we want the reliability of the positive test to be virtually perfect, to rule in disease; therefore the test has to be very specific [32].

Defining a screening or diagnostic test involves both the technique and the cut-off point used. For a given test and a given prevalence of abnormality in those tested, if the cut-off point is made more extreme (e.g. a higher concentration of αfetoprotein in serum), the sensitivity will fall (as more affected subjects will give a normal test) but the specificity will rise (as fewer unaffected subjects test positive). The optimum balance between sensitivity and specificity depends on the consequences of each.

ROC curves

One way to describe the effects of changing the cut-off is by a *receiver operating characteristic* (ROC) curve. These curves, or plots, were developed in the context of radar signal detection to separate signals from noise, and the radio terminology has persisted. An ROC plot is a graph of the sensitivity of the test against 1 – specificity, i.e. the true-positive rate against the false-positive rate. **Exhibit** 5.8 shows data for different cut-off points of measurements of prostate-specific antigen (PSA) in blood, giving the sensitivity and specificity for the detection of prostate cancer in men. Two ROC curves shown, one relating to all prostate cancer and the other to only aggressive cancers, with a pathological appearance giving a Gleason score of 7 or higher. A completely uninformative test would show a straight line from lower left to top right. A test that is better than random classification shows a curve above the line; the area under that curve indicates the probability that a subject with the disease has a higher measurement than a subject without the disease. The ideal test will produce a curve going close to the top left-hand part of the graph. These curves show that the test is clearly better than chance, and is better for more aggressive cancers. The graphs show the sensitivity obtained when an acceptable level of specificity is chosen, and vice versa. The ideal cut-off will depend on the consequences of false-positive and false-negative results. For example, if it is acceptable for 15 per cent of men to receive a positive-test result and go through the necessary further tests which that implies, the test at 0.15 for 1 – specificity will detect about 50 per cent of men with aggressive prostate cancer, while 50 per cent will have lower levels and be reported as ‘normal’. It is clear that if the cut-off was set to detect a high proportion of aggressive prostate cancers, such as 80 per cent, many men (nearly 40 per cent of those tested) would have
p. 144 false positive results. The authors of this ↳ study commented that the results challenge the assumption that there is a ‘normal’ PSA level, and that they show that there is no clearly defined cut-off point [33].



PSA (ng/ml)	All cancer		Gleason ≥ 7	
	Sensitivity	1 – Specificity	Sensitivity	1 – Specificity
1.1	0.83	0.57	0.93	0.61
1.6	0.67	0.38	0.85	0.42
2.1	0.55	0.27	0.75	0.31
2.6	0.45	0.19	0.66	0.23
3.1	0.37	0.15	0.54	0.18
4.1	0.28	0.08	0.43	0.11
6.1	0.06	0.03	0.15	0.03
8.1	0.03	0.01	0.06	0.01
10.1	0.01	0.01	0.02	0.01

Ex. 5.8. ROC plot (receiver operating characteristic): plot of sensitivity versus 1 specificity, and data for various cut-off points of PSA concentration, for PSA blood level compared with presence of any prostate cancer, and presence of aggressive prostate cancer (Gleason grade 7 or higher). Reproduced with permission from Thompson *et al*., JAMA, 294, 66–70; copyright 2005, American Medical Association, all rights reserved [33]

Curves can be compared by calculating the area under the curve (AUC) as a proportion of the total area of the graph; a ‘better’ test gives a higher AUC. Here the AUC was 0.78 for aggressive prostate cancer, but lower (0.68) for all cancer. Extensive further mathematical analysis can be done [34].

The ROC plot is valuable for comparing different tests. For example, to compare the new technique of digital mammography with the established method of film mammography (the difference is analogous to the difference between a digital camera and one using film), nearly 50 000 women in the USA and Canada were screened by both methods, in random order, and each set of images was assessed independently by two radiologists, using a seven-point scale for reporting the suspicion of malignancy. The analysis used ROC curves and the assessment of the AUC as the main statistic. While the two methods were similar in performance p. 145 overall, digital mammography was ↳ superior in women under the age of 50, as shown by the area under the ROC curve being greater. It was also better for women with dense breast tissue, and for women before or close to menopause [35].

Other measures of consistency of observations

The assessment of the consistency or inconsistency of observations is a large topic with extensive literature. The kappa method is described here because it is mathematically simple, and the effects of non-differential misclassification can be adequately explained by its use. However, there are many other methods, which are computationally more complex although similar in concept. One widely used measure is the *intra-class correlation coefficient*, which is derived from an analysis of variance. It is the ratio of the variance attributed to the different observation methods (e.g. two or more observers or episodes of observation) to the total variance, which is this plus the intra-subject variance. Like kappa, the range is from 1 to zero. A value of 1 means that all the variance in the data is equal to the variation between subjects and none of it is due to variation between observers, which means that there is no misclassification. Fuller discussion of these and other methods is given by, for example, Armstrong *et al*. [36].

Effects of error (non-differential misclassification) on study results

Clearly, misclassification of the exposure or the outcome will affect the results of a study. Error (non-differential misclassification) is misclassification of the same degree and direction in the different groups being compared. It will lead to a dilution of the effect, i.e. the observed odds ratio or other measure of association will be closer to the null value than is the true situation. (There are some exceptions to this, which will be mentioned later.) The extent of this bias may be considerable. We will describe the effects of non-differential misclassification on odds ratio estimates in a simple study design.

Approximate result based on kappa:

Two approximate results are very useful. For a simple study of any design with a 2 × 2 format, comparing diseased and non-diseased subjects, and exposed or unexposed subjects, for non-differential misclassification there is an approximate relationship between the observed and the true odds ratios and the value of kappa [37]:

$$\mathrm{OR}_\mathrm{O} = \kappa\,(\mathrm{OR}_\mathrm{T} - 1) + 1$$

where OR_O is the observed odds ratio and OR_T is the true odds ratio. The extremes of this formula show that if p. 146 kappa is 1.0, perfect assessment, the ↳ observed OR equals the true OR, and if kappa is zero, i.e. the measure of exposure is no better than chance, the observed OR is 1.0 irrespective of the true association.

For example, if the true odds ratio in a case–control study is 3.0, and the mea sure of exposure used has a kappa value of 0.8 compared with the true measure of exposure, the observed odds ratio will be about 0.8 (3.0 − 1) +1 =2.6.

If we know kappa (from a study of consistency of data), we can adjust an observed odds ratio to give an estimate of the true odds ratio. The formula is

$$OR_T = (\kappa + OR_O - 1)/\kappa.$$

The study shown in Ex. 5.6 was related to a case–control study showing an odds ratio of 1.51 between having had freckles and developing melanoma. Using the kappa value of 0.83 gives an estimated true odds ratio of (0.83 + 1.51 − 1)/0.83 = 1.61, i.e. modestly increased.

Approximate result based on continuous exposure measures

Information on the reproducibility or validity of an exposure measurement is often expressed as the correlation between measurements over the range of the variable. For a continuous measure, such as caloric intake, number of cigarettes smoked per day, or level of blood pressure, the extent of non-differential misclassification can be expressed as the *validity coefficient v*, which is the correlation between the observed measure of the exposure and its true value [36]. The relationship between the observed odds ratio OR_o related to a unit change in the measured variable and the true odds ratio OR_T is given by the square of this validity coefficient:

$$OR_O = OR_T^{\nu^2}$$

or

$$OR_O = \exp\,(\nu^2 \ln OR_T)$$

and

$$OR_T = \exp\left(\ln OR_O/\nu^2\right)$$

Suppose that a study shows an odds ratio of 1.50 with a unit increase in obesity, as measured in a field survey, and the correlation between that measurement of obesity and the true value, assessed by comparing the field measurement with an ideal measurement on an adequate sample of subjects, is 0.7. Then the true odds ratio is exp(ln 1.5/0.49)= 2.29.

p. 147 The validity coefficient will often be unknown, as the true value of the quantity may be difficult or impossible to measure (e.g. exposures in the past). Often all that is available is information from repeated measurements. The correlation between two measures of an exposure is the *reliability coefficient r*. Under ideal conditions this is equal to the square of the validity coefficient, and so *r* can be substituted for v^2 in the above equations. However, in practice such a measure should be assumed to be the upper limit, or most optimistic estimate, of validity.

For example, suppose a study assessing a relationship of adult disease to alcohol consumption in the teenage years yields an odds ratio of 2.0. There is no method of measuring the true value of this variable, but repeated measures using the same questionnaire will give a reliability coefficient. If this were 0.80, the revised estimate of the true odds ratio would be exp(ln 2.0/0.80) =2.38.

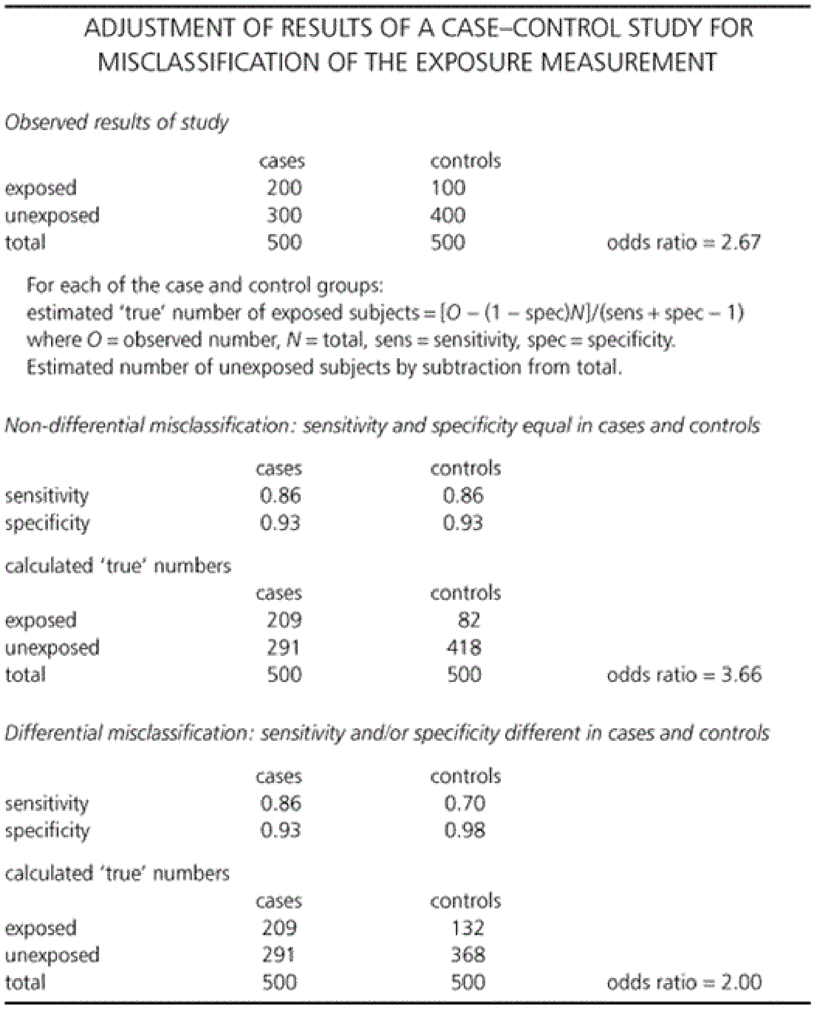
Use of estimates of sensitivity and specificity to calculate the effects of non-differential misclassification

We have seen how the effects of non-differential misclassification on study results can be estimated using an overall measure of agreement, kappa. Information of the sensitivity and specificity of the measures can also be used to estimate the effects of the misclassification. We will discuss this in the context of case–control studies, although the main results apply to all study designs.

The odds ratio seen in a study is related to the true odds ratio in a way that depends on sensitivity, specificity, and the prevalence of exposure prevalence in the control group. The arithmetic required to adjust observed results using known values of sensitivity and specificity is simple, if rather tedious. The sensitivity and specificity estimates are applied to the observed data to produce estimated ‘true’ data giving the numbers of exposed and unexposed cases and controls, and then the ‘true’ odds ratio is calculated from these data.

Non-differential misclassification means that sensitivity and specificity are known or assumed to be the same in cases and controls. Non-differential misclassification will move the observed odds ratio closer to the null value, i.e. it will dilute the effect, but the observed association will be in the correct direction as long as the sum of sensitivity and specificity is greater than 1.

Exhibit 5.9 shows a hypothetical example of a case–control study where the observed odds ratio is 2.67. If the sensitivity of the exposure assessment is 0.86 and the specificity is 0.93, and these figures apply to both cases and controls (i.e. the misclassification is non-differential), the estimated true odds ratio must be higher than the observed ratio. The calculations show that it is considerably higher (3.66). We shall discuss the lower part of Ex. 5.9 later in this chapter.



ADJUSTMENT OF RESULTS OF A CASE–CONTROL STUDY FOR MISCLASSIFICATION OF THE EXPOSURE MEASUREMENT

Observed results of study

	cases	controls	
exposed	200	100	
unexposed	300	400	
total	500	500	odds ratio = 2.67

For each of the case and control groups:
estimated 'true' number of exposed subjects = [*O* − (1 − spec)*N*]/(sens + spec − 1)
where *O* = observed number, *N* = total, sens = sensitivity, spec = specificity.
Estimated number of unexposed subjects by subtraction from total.

Non-differential misclassification: sensitivity and specificity equal in cases and controls

	cases	controls
sensitivity	0.86	0.86
specificity	0.93	0.93

calculated 'true' numbers

	cases	controls	
exposed	209	82	
unexposed	291	418	
total	500	500	odds ratio = 3.66

Differential misclassification: sensitivity and/or specificity different in cases and controls

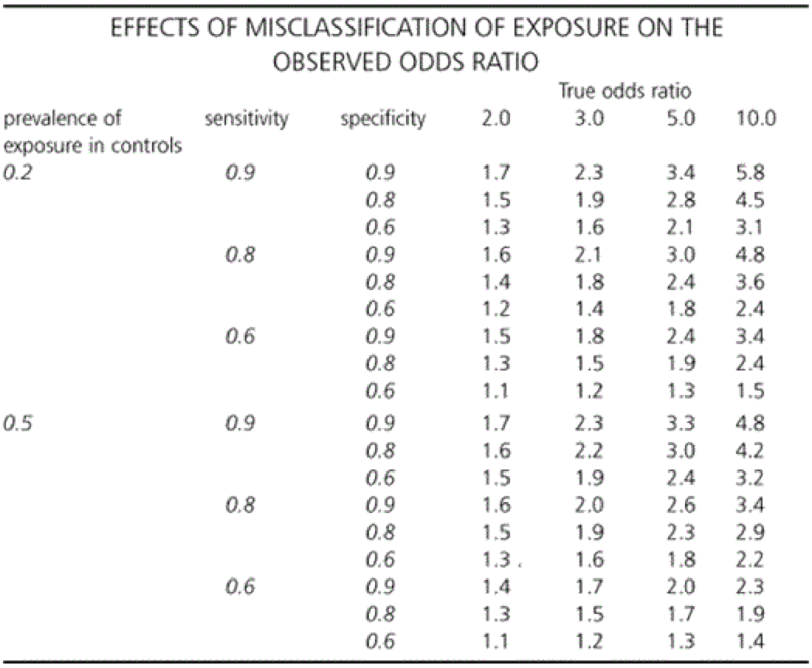
	cases	controls
sensitivity	0.86	0.70
specificity	0.93	0.98

calculated 'true' numbers

	cases	controls	
exposed	209	132	
unexposed	291	368	
total	500	500	odds ratio = 2.00

Ex. 5.9. Calculation of estimated 'true' results, adjusted for misclassification: hypothetical example of observed results of a case–control study with independent information on the sensitivity and specificity of the exposure measure

p. 148 A table illustrating the relationship between the true and the observed odds ratios with non-differential
misclassification is shown in **Ex** 5.10. This shows that the effect on the odds ratio can be quite considerable. As
mentioned, the extent of the reduction towards the null of the observed odds ratio depends on the sensitivity
and specificity values, and also on the prevalence of exposure. For example, if sensitivity and specificity are
both 90 per cent, which would be regarded as quite good in most circumstances, and the prevalence of
p. 149 exposure ↳ in the control group is 20 per cent, a true association with an odds ratio of 5.0 would relate to an
observed odds ratio of only 3.4.



EFFECTS OF MISCLASSIFICATION OF EXPOSURE ON THE OBSERVED ODDS RATIO

			True odds ratio			
prevalence of exposure in controls	sensitivity	specificity	2.0	3.0	5.0	10.0
0.2	*0.9*	*0.9*	1.7	2.3	3.4	5.8
		0.8	1.5	1.9	2.8	4.5
		0.6	1.3	1.6	2.1	3.1
	0.8	*0.9*	1.6	2.1	3.0	4.8
		0.8	1.4	1.8	2.4	3.6
		0.6	1.2	1.4	1.8	2.4
	0.6	*0.9*	1.5	1.8	2.4	3.4
		0.8	1.3	1.5	1.9	2.4
		0.6	1.1	1.2	1.3	1.5
0.5	*0.9*	*0.9*	1.7	2.3	3.3	4.8
		0.8	1.6	2.2	3.0	4.2
		0.6	1.5	1.9	2.4	3.2
	0.8	*0.9*	1.6	2.0	2.6	3.4
		0.8	1.5	1.9	2.3	2.9
		0.6	1.3 .	1.6	1.8	2.2
	0.6	*0.9*	1.4	1.7	2.0	2.3
		0.8	1.3	1.5	1.7	1.9
		0.6	1.1	1.2	1.3	1.4

Ex. 5.10. Observed odds ratios in a case–control study with non-differential classification, for given true odds ratio, sensitivity and specificity of exposure assessment, and prevalence of exposure in the control group

These calculations should be used only as a general guide. In most situations the information on the sensitivity and specificity of the measurements of exposure or outcome in studies is limited, and is often obtained from other studies. The estimates of sensitivity and specificity will themselves have sampling errors dependent on the numbers of observations used to produce them. Therefore, although it is possible to use tables such as Ex. 5.10 or calculations to estimate the odds ratio from an observed odds ratio, such estimates must be treated as approximate. Mathematical adjustment of the observed results of a study is usually used in published studies, if at all, only as a discussion point for interpretation.

Further effects of non-differential misclassification

We will mention some further issues with misclassification; but these are of less general relevance.

p. 150

Situations in which non-differential misclassification can bias results away from the null: reversal of effect with extreme misclassification

As stated already, non-differential misclassification will almost always reduce the estimate of association towards the null value. There are some extreme situations where this generalization does not hold, which will be described briefly. In a case–control study where exposure is being assessed, if there is no misclassification, so that sensitivity and specificity are both 100 per cent, there is no bias; the observed odds ratio will equal the true odds ratio. As the sensitivity or the specificity of the exposure assessment is reduced, the odds ratio shifts towards the null value until, when the sum of sensitivity and specificity is equal to 1.0, the observed odds ratio will be 1.0. This will happen for any value of the true odds ratio. This situation is the equivalent of the exposure assessment being no better than labelling cases and controls as being exposed or unexposed by a random process. If the sum of sensitivity and specificity is less than 1 (i.e. sensitivity and specificity average less than 50 per cent), the assessment of exposure is worse than chance assignment, and the odds ratio can show an association in the opposite direction to the true association. Therefore very severe misclassification can reverse the direction of an association. This situation also applies to non-differential misclassification of the exposure in a cohort study.

Effects of non-differential misclassification of the outcome in cohort studies

In a cohort design, subjects are selected on their exposure and the outcome is assessed. The effects of non-differential misclassification of the outcome depend differently on sensitivity and specificity. A reduction in sensitivity from the ideal of 100 per cent will have no effect of the relative risk estimate, but will reduce the observed risk difference. A reduction in specificity will reduce both the relative risk and risk difference estimates [38]. Even at very low levels of sensitivity and specificity the direction of the true association will not be reversed.

Effects if there are more than two categories

If the exposure is in more than two categories, the effects of non-differential misclassification may be more complex. If there are categories of unexposed, moderately exposed, and highly exposed, and the true situation is that the risk increases across that gradient, the effect of misclassification will be to reduce the observed association in the highly exposed group (as it will contain more individuals who are not actually highly exposed), but it could either increase or decrease the observed risk in the moderately exposed category. Thus the effects of non-differential misclassification for exposures in more than two ↳ categories may be complex, and can sometimes result in a bias of the trend estimate away from the null. p. 151

Non-independent errors

A further situation in which non-differential error can produce biases away from the null is where the errors in the ascertainment of the exposure and of the outcome are not independent. For example, in a survey using several interviewers and subjective data, interviewer differences in assessing exposure and outcome may produce related effects. Such situations need special caution, but the effects will be specific to the situation.

Misclassification of confounders

Confounding will be discussed in Chapter 6. The ability to adjust for the affects of confounding will depend on the accuracy of measuring the confounder. If the confounder is only approximately measured, the ability to control its true confounding affect is limited. It follows that the non-differential misclassification of a confounding variable will reduce the degree to which the confounding can be controlled. An example will be given in Chapter 6.

Effects on the numbers of subjects needed in a study

A useful application of this, and the situation in which the calculation of observed values from assumed true values is useful, is in the estimation of the size of a projected study. As will be discussed in Chapter 7, this estimation depends on the odds ratio that is assumed to apply. As the observed odds ratio will be closer to the null than the true ratio because of misclassification, it is prudent to take this into account by using the projected observed odds ratio when calculating sample sizes.

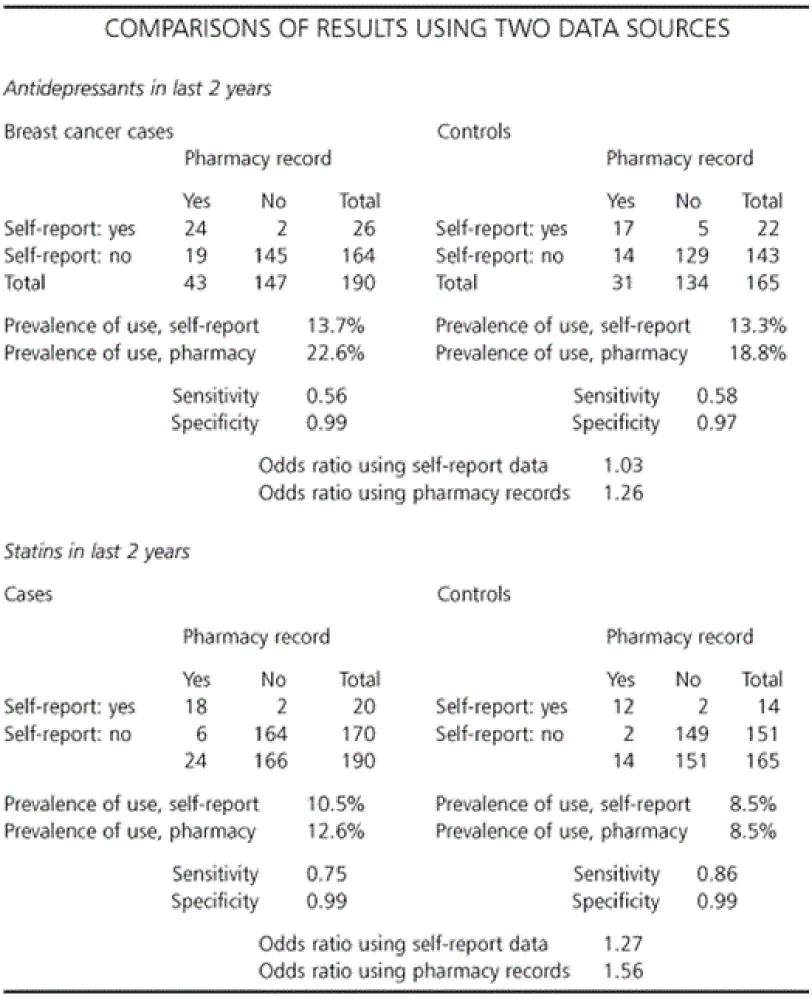
The situations are of interest. However, it is still a reasonable assumption that in the great majority of studies, imperfect assessment of exposure and outcome will show itself in decreases in both sensitivity and specificity, and the inaccuracy of assessments is likely to reduce the observed risk ratio and risk difference estimates towards the null.

Differential misclassification

Differential misclassification can produce bias in either direction, and of almost any magnitude; the observed measure of association can be higher or lower than the true result. In case–control studies, odds ratio and relative risk estimates will tend to be increased (further from the null) if the sensitivity is higher and/or the specificity is lower for the measure of exposure assessment in cases than for controls.

p. 152 The misclassification being 'differential' means that the sensitivity and/or specificity is different for cases and controls. In the example in the lower part of Ex. 5.9, sensitivity is 0.86 for cases but 0.70 for controls, and specificity is 0.93 for cases but 0.98 for controls, and the 'true' odds ratio is 2.00, lower than the observed value of 2.67. Therefore in this example, differential misclassification causes the observed odds ratio to be
p. 153 exaggerated. Of course, examples can be ↳ produced where differential effects have a wide range of outcomes. Tables similar to that in Ex. 5.10 can be generated by the same calculations, but with different values of sensitivity and/or specificity for cases and for controls.

As a real example, Boudreau *et al*. [39] performed a case–control study of breast cancer in Washington State in members of a health plan with its own pharmacies. The primary source of information on the use of medical drugs was interviews, but the interview data could be checked against pharmacy records, which were accepted as a gold standard. As shown in **Ex.** 5.11, the prevalence of use of antidepressants was considerably underestimated in the interview, with only 13.7 per cent of breast cancer patients reporting use in the previous 2 years, compared with 22.6 per cent reported by pharmacy records. However, this considerable inaccuracy was largely non-differential; the prevalence of use was under-reported by both breast cancer cases and controls. The table shows that the sensitivity of self-report was 56 per cent in cases and 58 per cent in controls, while the specificity was 99 per cent in cases and 97 per cent in controls. The odds ratio of the association between antidepressant use and breast cancer based on self-reported data is 1.03, while the odds ratio using pharmacy records is 1.26.



COMPARISONS OF RESULTS USING TWO DATA SOURCES

Antidepressants in last 2 years

Breast cancer cases				Controls			
	Pharmacy record				Pharmacy record		
	Yes	No	Total		Yes	No	Total
Self-report: yes	24	2	26	Self-report: yes	17	5	22
Self-report: no	19	145	164	Self-report: no	14	129	143
Total	43	147	190	Total	31	134	165

Prevalence of use, self-report	13.7%	Prevalence of use, self-report	13.3%
Prevalence of use, pharmacy	22.6%	Prevalence of use, pharmacy	18.8%
Sensitivity	0.56	Sensitivity	0.58
Specificity	0.99	Specificity	0.97

Odds ratio using self-report data	1.03
Odds ratio using pharmacy records	1.26

Statins in last 2 years

Cases				Controls			
	Pharmacy record				Pharmacy record		
	Yes	No	Total		Yes	No	Total
Self-report: yes	18	2	20	Self-report: yes	12	2	14
Self-report: no	6	164	170	Self-report: no	2	149	151
	24	166	190		14	151	165

Prevalence of use, self-report	10.5%	Prevalence of use, self-report	8.5%
Prevalence of use, pharmacy	12.6%	Prevalence of use, pharmacy	8.5%
Sensitivity	0.75	Sensitivity	0.86
Specificity	0.99	Specificity	0.99

Odds ratio using self-report data	1.27
Odds ratio using pharmacy records	1.56

Ex. 5.11. Results based on two data sources: results for the association between breast cancer and use of prescribed drugs in the previous 2 years based on self-report and on pharmacy records that are regarded as more valid. From Boudreau *et al*. [39]

With regard to the past use of statins, the accuracy of the information appears to be better. The prevalence of use in cases based on self-report was 10.5 per cent compared with 12.6 per cent based on pharmacy records, and in the control group was 8.5 per cent by both sources. The difference between self-report and pharmacy records is considerably smaller than that for antidepressants, but is more differential. The sensitivity of self-report was 75 per cent in cases but 86 per cent in controls, although the specificity was 99 per cent in each group. The self-reported data give the association between past use of statins and breast cancer as an odds ratio of 1.27, again below the odds ratio of 1.56 that is given by the pharmacy records.

Self-test questions (answers on p. 498)

Q5.1 Which should be considered first in interpretation of a study: chance variation, observation bias, or confounding?

Q5.2 What is the essential distinction between error and bias?

Q5.3 What is the effect of error (non-differential error) on the results of a study?

Q5.4 What is the effect of observation bias on the results of a study?

Q5.5 What do the terms single-blind, double-blind, and triple-blind imply?

Q5.6 What are the main practical methods available to reduce bias and error?

p. 154 Q5.7 ↳ In a study in which a yes/no question was asked on two occasions, 60 per cent of subjects answered 'yes' on each occasion, 20 per cent answered 'no' on each occasion, 10 per cent answered 'yes' on the first test and 'no' on the second, and 10 per cent answered 'no', then 'yes'. Calculate:

(a) The proportion showing agreement.

(b) The expected agreement by chance.

(c) Kappa.

Q5.8 If the study in Q5.7 showed an odds ratio of 1.8 for the association between the factor assessed and outcome, what would the true odds ratio be?

Q5.9 For general purposes, how would kappa values of 0.9, 0.7, 0.5, and 0.2 be interpreted?

Q5.10 A screening test applied to 1000 subjects identified 100 of them as positive. On further investigation, 50 of these were found to have the disease being tested for. In subsequent follow up, 10 of the subjects who tested negative developed the disease within a reasonably short time. Calculate the sensitivity, specificity, and predictive value positive of the test.

Q5.11 A survey using a simple portable method of measuring haemoglobin level shows an odds ratio of 1.2 for the association of disease with a unit change in haemoglobin level. The correlation between this field mea surement and the best laboratory method is 0.9. What is the estimated true odds ratio?

Q5.12 In a case–control study with a dichotomous exposure variable, 50 per cent of cases and 25 per cent of controls are exposed. For the exposure measurement, the sensitivity is 0.80 and specificity 0.95. Assuming non-differential misclassification, what is the true odds ratio?

References

1. Brenner H, Arndt V, Stegmaier C, Ziegler H, Rothenbacher D. Is *Helicobacter pylori* infection a necessary condition for noncardia gastric cancer? *Am J Epidemiol* 2004; **159**: 252–258. 10.1093/aje/kwh039
Crossref

2. Leclerc A, Martinez CM, Gerin M, Luce D, Brugere J. Sinonasal cancer and wood dust exposure: results from a case–control study. *Am J Epidemiol* 1994; **140**: 340–349.
WorldCat

3. Fikree FF, Gray RH, Shah F. Can men be trusted? A comparison of pregnancy histories reported by husbands and wives. *Am J Epidemiol* 1993; **138**: 237–242.
WorldCat

4. Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis. III: The lack of support for a genetic hypothesis. *J Chronic Dis* 1969; **22**: 217–222. 10.1016/0021-9681(69)90015-0
WorldCat Crossref

5. Carter CO, David PA, Laurence KM. A family study of major central nervous system malformations in South Wales. *J Med Genet* 1968; **5**: 81–106. 10.1136/jmg.5.2.81
WorldCat Crossref

p. 155 6. Mandel EM, Rockette HE, Bluestone CD, Paradise JL, Nozza RJ. Efficacy of amoxicillin with and without decongestant-antihistamine for otitis media with effusion in children. Results of a double-blind, randomized trial. *N Engl J Med* 1987; **316**: 432–437. 10.1056/NEJM198702193160803
WorldCat Crossref

7. Cantekin EI, McGuire TW, Griffith TL. Antimicrobial therapy for otitis media with effusion (‘secretory’ otitis media). *JAMA* 1991; **266**: 3309–3317. 10.1001/jama.266.23.3309
WorldCat Crossref

8. Rennie D. The Cantekin affair. *JAMA* 1991; **266**: 3333–3337. 10.1001/jama.266.23.3333
WorldCat Crossref

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

10. Seaton ED, Charakida A, Mouser PE, Grace I, Clement RM, Chu AC. Pulsed-dye laser treatment for inflammatory acne vulgaris: randomised controlled trial. *Lancet* 2003; **362**: 1347–1352. 10.1016/S0140-6736(03)14629-6
WorldCat Crossref

11. **Medical Research Council**. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948; **ii**: 769–782.
WorldCat

12. Tucker JS, Hall MH, Howie PW, *et al*. Should obstetricians see women with normal pregnancies? A multicentre randomised controlled trial of routine antenatal care by general practitioners and midwives compared with shared care led by obstetricians. *BMJ* 1996; **312**: 554–559. 10.1136/bmj.312.7030.554
WorldCat Crossref

13. Weatherley-Jones E, Nicholl JP, Thomas KJ, *et al*. A randomised, controlled, triple-blind trial of the efficacy of homeopathic treatment for chronic fatigue syndrome. *J Psychosom Res* 2004; **56**: 189–197. 10.1016/S0022-3999(03)00377-5
WorldCat Crossref

14. Burns KE, Chu MW, Novick RJ, *et al*. Perioperative *N*-acetylcysteine to prevent renal dysfunction in high-risk patients

undergoing CABG surgery: a randomized controlled trial. *JAMA* 2005; **294**: 342–350. 10.1001/jama.294.3.342
Crossref

15. Royal College of General Practitioners. *Oral Contraceptives and Health: An Interim Report from the Oral Contraception Study of the Royal College of General Practitioners*. New York: Pitman Medical, 1974.
Google Scholar Google Preview WorldCat COPAC

16. Vessey M, Doll R, Peto R, Johnson B, Wiggins P. A long-term follow-up study of women using different methods of contraception-an interim report. *J Biomed Sci* 1976; **8**: 373–427.
WorldCat

17. Vessey MP. Some methodological problems in the investigation of rare adverse reactions to oral contraceptives. *Am J Epidemiol* 1971; **94**: 202–209.
WorldCat

18. 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

19. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83 000 women with breast cancer from 16 countries. *Lancet* 2004; **363**: 1007–1016. 10.1016/S0140-6736(04)15835-2
WorldCat Crossref

20. 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

21. Johannes CB. Interviewer effects in a cohort study: results from the Massachusetts Women's Health Study. *Am J Epidemiol* 1997; **146**: 429–438. 10.1093/oxfordjournals.aje.a009296
WorldCat Crossref

22. Paul C, Skegg DCG, Spears GFS, Kaldor JM. Oral contraceptives and breast cancer: a national study. *BMJ* 1986; **293**: 723–726. 10.1136/bmj.293.6549.723
WorldCat Crossref

23. Grayson DA, O'Toole BI, Marshall RP, *et al*. Interviewer effects on epidemiologic diagnosis of posttraumatic stress disorder. *Am J Epidemiol* 1996; **144**: 589–597. 10.1093/oxfordjournals.aje.a008969
WorldCat Crossref

p. 156 24. Westerdahl J, Anderson H, Olsson H, Ingvar C. Reproducibility of a self-administered questionnaire for assessment of melanoma risk. *Int J Epidemiol* 1996; **25**: 245–251. 10.1093/ije/25.2.245
WorldCat Crossref

25. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**: 37–46. 10.1177/001316446002000104
WorldCat Crossref

26. Fleiss JL. *Statistical Methods for Rates and Proportions* (2nd edn). New York: John Wiley, 1981.
Google Scholar Google Preview WorldCat COPAC

27. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. The evidence for accuracy of recall. *Am J Epidemiol* 1989; **129**: 233–248.
WorldCat

28. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine* (2nd edn). Boston,

MA: Little Brown, 1991.
Google Scholar Google Preview WorldCat COPAC

29. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491–1494. 10.1136/bmj.304.6840.1491
WorldCat Crossref

30. Aitken JF, Youl P, Janda M, *et al*. Validity of self-reported skin screening histories. *Am J Epidemiol* 2004; **159**: 1098–1105. 10.1093/aje/kwh143
WorldCat Crossref

31. Wald NJ, Cuckle HS, Boreham J, *et al*. Antenatal screening in Oxford for fetal neural tube defects. *Br J Obstet Gynaecol* 1979; **86**: 91–100. 10.1111/j.1471-0528.1979.tb10574.x
WorldCat Crossref

32. Institute for Clinical Evaluative Sciences. The jargon decoder: interpreting diagnostic tests. If you expect me to be sensitive, then don't ask me to be so specific. *Informed* 1998; **4**: 1–2.
WorldCat

33. Thompson IM, Ankerst DP, Chi C, *et al*. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/mL or lower. *JAMA* 2005; **294**: 66–70. 10.1001/jama.294.1.66
WorldCat Crossref

34. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; **39**: 561–577.
WorldCat

35. Pisano ED, Gatsonis C, Hendrick E, *et al*. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005; **353**: 1773–1783. 10.1056/NEJMoa052911
WorldCat Crossref

36. Armstrong BK, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*. Oxford: Oxford University Press, 1992.
Google Scholar Google Preview WorldCat COPAC

37. Thompson WD, Walter SD. Variance and dissent. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988; **41**: 949–958. 10.1016/0895-4356(88)90031-5
WorldCat Crossref

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

39. Boudreau DM, Daling JR, Malone KE, Gardner JS, Blough DK, Heckbert SR. A validation study of patient interview data and pharmacy records for antihypertensive, statin, and antidepressant medication use among older women. *Am J Epidemiol* 2004; **159**: 308–317. 10.1093/aje/kwh038
WorldCat Crossref