Chapter 2
Measurement of exposures and outcomes

2.1 Introduction

Most epidemiological research involves the study of the relationship of one type of event or characteristic to another. Consider the following questions as examples:

* *Does alcohol intake increase the risk of lung cancer?*

  Alcohol → lung cancer
  (exposure) (outcome)

* *Does hepatitis B vaccination protect against liver cancer?*

  Hepatitis B vaccine → liver cancer
  (exposure) (outcome)

In these relationships, we assume that one event—exposure—affects the other—outcome.

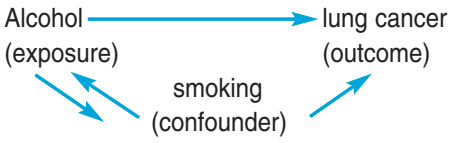
The *exposure* of interest may be associated with either an increased or a decreased occurrence of disease or other specified health outcome, and may relate to the environment (e.g., air pollution, indoor radon), lifestyle (e.g., smoking habits, diet), or inborn or inherited characteristics (e.g., blood group A, fair skin). The term risk factor is often used to describe an exposure variable.

The *outcome* of a study is a broad term for any defined disease, state of health, health-related event or death. In some studies, there may be multiple outcomes.

The exposures and outcomes of interest are specific to study hypotheses and should always be clearly defined before the study starts. The exposure of interest in one study may be the outcome in another. For example, smoking is clearly the exposure of interest in a study that examines whether smokers are more likely to develop lung cancer than non-smokers, but would be the outcome in a study examining the effectiveness of an anti-smoking intervention programme in reducing the frequency of smoking in a certain population.

In most instances, it is not sufficient to collect information only on the exposure and outcome of interest. This is because their relationship may be mixed up with the effect of another exposure on the same outcome, the two exposures being correlated. This phenomenon is known as confounding. Consider again the relationship between alcohol intake and lung cancer.

* *Does alcohol intake increase the risk of lung cancer?*



Suppose that a researcher observes that lung cancer occurs more often in people who drink alcohol than in those who do not. It would not be possible to conclude from this observation that exposure to alcohol increases the probability of developing lung cancer, unless the researcher can show that the observed relationship cannot be due to the fact that those who drink alcohol smoke more heavily than non-drinkers. In this example, smoking is acting as a *confounder*. Confounding can be dealt with when designing studies or when analysing the results provided that the relevant data have been collected. These issues are discussed in detail in Chapters 13 and 14.

Thus, most epidemiological studies must collect information on three types of variable:

(1) the primary exposure(s) of interest,
(2) other exposure(s) that may influence the outcome (potential confounders), and
(3) the outcome(s).

It is impossible to select appropriate measurements for a particular investigation unless a specific and detailed statement of research objectives has been made. Without such a statement, information on key variables may be inadequate or missing.

This chapter discusses different ways of collecting data on exposures and outcomes.

2.2 Types of exposure

A wide range of exposures may be of interest in cancer epidemiology. These include genetic traits (e.g., blood group), demographic variables (e.g., sex, age, ethnicity, socioeconomic status), reproductive and sex-related variables, diet and body build, physical activity, smoking and alcohol habits, past medications (e.g., oral contraceptive use), environmental and occupational exposures, and infectious agents.

The characteristic of interest, the true exposure, may not be directly measurable, or it may be difficult or impossible to define. Socioeconomic status is an example of such an abstract concept. Epidemiologists commonly measure socioeconomic status using proxy variables such as occupation, income, education, and place of residence. Moreover, socioeconomic status is not *per se* a cause of disease, but rather an indicator of the level or probability of exposure to some underlying cause, which is often unknown.

2.3. Measurement of exposure

Data on the exposures of interest may be obtained through personal interviews (either face-to-face or by telephone), self-administered questionnaires, diaries of behaviour, reference to records, biological measurements and measurements in the environment. If a subject is too young, too ill, or dead, it is also common to obtain data from a proxy respondent, usually a member of their family.

The method chosen to collect data depends on many factors: the type of study; the type and detail of data required; availability of existing records collected for other purposes; lack of knowledge or poor recall of the exposure by subjects; sensitivity of the subjects to questioning about the exposure; frequency and level of the exposure, and their variability over time; availability of physical or chemical methods for measuring the exposure in the human body or in the environment; and the costs of the various possible methods. Often, more than one approach is used. Different components of the data often require different collection methods, and using several methods of data collection can help to validate data and to reduce error in measurement (see Section 2.6).

The information obtained should include details of the exact nature of the exposure, its amount or dose, and its distribution over time.

2.3.1 Nature of the exposure

The information collected should be as detailed as possible. For instance, it is better to enquire about different forms of tobacco smoking separately (cigarettes, pipes, cigars), rather than to enquire simply about 'smoking'. Questions on types of cigarette may also be asked to obtain information on their tar content. Enquiries should also be made about the route of exposure to the agent (for example, in a study of contraceptives and breast cancer, it is important to distinguish oral contraceptives from other types of contraceptive), as well as about any behaviour that may protect against exposure (for example, in an occupational study, it is important to ask about any behaviour that may have protected the workers from being exposed to hazards, such as use of protective clothing).

2.3.2 Dose

Exposure is seldom simply present or absent. Most exposures of interest are quantitative variables. Smokers can be classified according to the number of cigarettes smoked daily; industrial exposures by the extent of exposure (often achieved by classifying workers according to the duration of employment and type of job); infections by dose of agent or age at exposure; breast-feeding by duration; and psychological exposures by some arbitrary scale of severity. Thus the simple situation of two groups, one exposed and one unexposed, is rare, and the conclusions of a study are greatly strengthened where there is a trend of increasing disease incidence with increasing exposure—an exposure–response relationship.

Dose may be measured either as the total accumulated dose (cumulative exposure), for example, the total number of packets of cigarettes ever smoked, or as the dose or exposure rate, for example, the number of cigarettes smoked daily. Exposure rate is a measurement of dose per unit time.

It is important to realize that although measurements of dose are usually made in the subject's external environment (e.g., levels of environmental pollution), this is not the dose that matters in biological terms. The biologically effective dose is the amount of the external agent or its active metabolite that affects cellular mechanisms in the target organs. The biologically effective dose cannot usually be measured directly, but it may be possible to obtain an estimate, an example being the measurement in humans of DNA adducts with nitrosamines or aflatoxins. Nevertheless, such measurements have their limitations: for instance, they may be useful markers of current or recent, but not of past, exposure (see Section 2.4.4).

2.3.3 Time

As far as possible, each exposure should be characterized as to when it began, when it ended (if at all), and how it was distributed during the intervening period (was it periodic or continuous? did the dose vary over time?). Similar details should also be obtained for any behaviour that may protect against the exposure.

There is thought to be a restricted period, the critical time window, during which the exposure could have caused cancer. Unfortunately, the beginning and end of this critical time window are not known, and its length is likely to vary between individuals. Collecting data on the timing of exposure allows the possible extent of this window to be estimated. Analyses may include examination of the effects of time since first exposure and time since last exposure.

Pattern of exposure may also be important. Exposure that occurs periodically in intense bursts may have a different effect from a similar total amount of exposure that occurs continuously at low intensity (e.g., constant versus intermittent exposure to sunlight; chronic exposure to low levels of ionizing radiation versus acute exposure to high levels).

2.4 Sources of exposure data

2.4.1 Questionnaires

Questionnaires are used to collect exposure data in epidemiological studies by putting the same set of questions to each study participant in a standardized form. Questionnaires can be self-administered or may be administered by an interviewer.

The aim of a research questionnaire is to obtain, with minimal error, measurements of the exposure variables of interest for the study. Thus, the questions to be included in a questionnaire should relate directly to the objectives of the study. Some basic principles that should be taken into account when designing a questionnaire are discussed in Appendix 2.1. To

ensure that the questions are properly understood and will elicit appropriate answers, questionnaires should be pre-tested on a sample of subjects from the population to be studied.

Self-administered questionnaires

Self-administered questionnaires are distributed to study subjects who are asked to complete them. They can be delivered and returned either personally or by mail if this is feasible and more convenient. Such questionnaires are particularly appropriate when small amounts of reasonably simple data are required, or for documenting sensitive or socially undesirable behaviour. They are one of the cheapest ways of collecting information, but have the limitation that they can be used only in literate populations. The investigator also has relatively little control on the quality of the data collected.

Personal interviews (interviewer-administered questionnaires)

Using an interviewer to administer a questionnaire may reduce error by increasing the subjects' participation and motivating them to respond adequately. Moreover, an interviewer may probe to obtain more complete data. However, interviewers may also increase error if they influence the subject's responses, either directly or indirectly.

As an interview is a conversation between interviewer and respondent, it is essential that a rapport is established right from the start. Interviewers should be selected taking into account the cultural norms of the study population, so that they will be trusted by the study subjects. As a simple example, in some societies, male interviewers will not be allowed to interview women. Cultural characteristics of interviewers may also influence the degree of participation of respondents, and/or the accuracy of the information they give. The respondent must feel that the interviewer understands him or her and that there are no barriers to communication.

For collecting large amounts of complex data, face-to-face interviews are clearly best. However, when subjects are widely dispersed and the questionnaire is relatively brief, interviewing by telephone may be a better approach. Of course, this is feasible only where the telephone is widely used, which is not always the case. Even in societies where there is widespread use of telephones, certain groups of people will be excluded from the study either because they do not have a telephone or because they do not like to provide personal information over the telephone.

Proxy or *surrogate respondents* are people who provide information on exposure in place of the study subjects themselves (index subjects). They are used in epidemiology when the index subjects are for any reason unable to provide the data required. Studies involving children normally also rely on proxy respondents. Proxy respondents usually provide less valid information than the index subjects; for instance, they often tend to under-enumerate occupational exposures and to report the index subject as having a job of higher status than is actually the case. Closeness to the

study subject is an important determinant of the quality of information obtained; in general, the most accurate information tends to come from spouses and, in the case of children, mothers.

2.4.2 Diaries

Diaries are detailed records of exposure kept by the subject. They are generally open-ended and take the form of a booklet in which the subject records each occurrence of a particular behaviour such as physical exercise, alcohol consumption, dietary intake, sexual activity, use of medication, etc. Diaries are assumed to be highly accurate in measuring current behaviour, because they do not rely on memory. They also allow more detailed information about the exposure to be collected than with a questionnaire. For example, foods can be weighed by the subject before being eaten.

The main limitation of diaries is that only current exposures can be measured. In addition, diaries generally demand more of subjects in terms of time and skill than other methods, so compliance may be a problem. Training of subjects in the skills needed to keep an accurate diary can be time-consuming for both subjects and investigators. Thus, diaries are rarely used in countries in which many people are illiterate.

2.4.3 Records

Data on the exposure of interest may be available from census, employment, medical (in- and out-patient), cancer registry, birth certification and death certification records. Typically, as the data have already been collected for purposes other than epidemiological research, the researcher has no control over what items were recorded, how questions were phrased, and so on. Records are also often produced by a large number of people with little uniform training. Moreover, the availability and quality of records in many countries tends to be poor.

Despite these limitations, the use of records has several advantages over other methods of data collection. Study costs are usually low, and the duration of the study is shorter because some or all of the data have already been collected. Records can also provide near-complete data on a well defined population, and information can be obtained without contacting the subjects or their relatives. Certain data items (for example, intake of medications or occupational exposures) may be recorded more accurately than information obtained in a personal interview; for instance, errors caused by poor recall or lack of knowledge of the exposure are eliminated.

Characteristics and limitations of some such routine data-collection systems are discussed in more detail in Section 2.9 and Chapter 11.

2.4.4 Biological measurements

In principle, the ideal approach to determining exposure involves measurements made directly on the human body or its products. Biological measurements will be more objective, in that they are independent both

of the subjects' perceptions and, where instrumental or laboratory methods are used, of the researcher. Biological measurements may also reflect more closely the biologically effective dose, i.e., the level of exposure that affects cells in the target organ(s).

Interest in the epidemiological application of measurements of exposure in the human body has recently been growing, with the development of increasingly refined laboratory techniques for measuring active metabolites of carcinogens and the products of their interaction with DNA or proteins (adducts). The term 'molecular epidemiology' has been coined to describe epidemiological approaches that incorporate a laboratory component.

An example of the successful application of molecular epidemiology is the measurement of aflatoxin in the human diet. Aflatoxin is produced by the mould *Aspergillus flavus*, which grows on stored foods such as groundnuts in tropical climates, in particular in eastern Asia and sub-Saharan Africa. Although experiments have shown that aflatoxin is a potent inducer of liver cancer in laboratory animals, most epidemiological research has been hampered by the difficulty of measuring the amount of aflatoxin consumed by humans. Recently, biological markers for estimating current or recent aflatoxin consumption have been established, involving measurement of metabolites of aflatoxin and DNA adducts in the urine. Such measurements were made in a study undertaken in Shanghai (Qian *et al.*, 1994), in which the incidence of liver cancer in approximately 18 000 Chinese men was related to urinary measurements of their exposure to aflatoxin. Results from this study have provided the most direct evidence that aflatoxin has an etiological role in human hepatocellular carcinogenesis. These biological markers are, however, not ideal, as they cannot measure past exposure, which may be crucial in studying the role of aflatoxin in liver cancer.

Laboratory assays have also been developed to ascertain exposure to infectious agents such as human papillomavirus (HPV) (Muñoz *et al.*, 1992b) and *Helicobacter pylori* (IARC, 1994a). These assays have helped to clarify the role of HPV infection in the etiology of cervical cancer, and that of *H. pylori* in stomach cancer.

The possibility of using laboratory measurements in an epidemiological study is determined mainly by the availability of a suitable test, its feasibility (e.g., availability of laboratory equipment) and the cost. Moreover, most laboratory measurements are limited in that they can assess only current exposures, while past exposure is generally more relevant in cancer epidemiology. Thus, laboratory measurements are particularly useful when they assess attributes that remain stable, for example, genetic traits. One way in which this limitation can be overcome is to use banks of biological specimens. Biological samples collected some time before the study subjects develop the outcome of interest can be analysed with the latest laboratory techniques. For instance, blood and urine samples may be collected from all individuals in a particular cohort at the time they enter the

study and an aliquot stored frozen. These samples can be re-analysed later when more sophisticated techniques become available.

2.4.5 Measurements in the environment

Measurements in the environment include those of agents in the air (e.g., air pollutants, dust), water (e.g., fluoride), soil (e.g., elements), foods (e.g., nutrient composition), etc. The samples may come from homes, workplaces, recreational sites, or the ambient environment in general. Such measurements are particularly useful when the subjects are unaware of the exposure (e.g., indoor radiation levels) or cannot recall it accurately.

The value of environmental measurements depends on the procedures used both for sampling and for analysis. Ideally, environmental agents should be assessed for each study subject throughout the etiologically relevant period, so as to reflect as accurately as possible personal attributes. For example, individual measurements of exposure to ionizing radiation can be made by each study subject wearing a film-badge throughout the study period and individual nutrient intake can be measured by analysing identical portions of all foods and beverages consumed by a subject during the study period. However, this approach is generally not feasible because of time and cost constraints, technical concerns and lack of subject compliance. Usually it is only possible to make measurements in a sample of study subjects at certain defined time points. The choice of the sample and the timing of the measurements is obviously of crucial importance to the validity of the measurements.

One limitation of environmental measurements is that they usually reflect only current exposure levels. In certain situations, it may be reasonable to assume that measurements made in the present environment are highly correlated with the exposure levels at etiologically relevant periods in the past. Records of previous exposure measurements may be available, but should be used with caution: such measurements were usually made for other purposes using methods that may now be considered inadequate. When no such measurements are available, proxy measures of past exposures may be used. For example, in a study of occupational exposures, information on 'type of job', 'year of employment' and 'duration of employment' may be used to classify workers according to exposure status. This information may be extracted from employment records or obtained through questionnaires.

2.5 Measurement of outcome

As for measurements of exposure, data on the outcome(s) of interest may be obtained from various sources. Regular questionnaires or telephone calls may be used to ascertain subjects' health status. Periodic personal interviews with clinical check-ups may be arranged, which may include biological measurements and any other appropriate diagnostic procedures (e.g., radiography, endoscopy, ultrasound, etc.). Alternatively,

information on the outcomes, and in particular on the occurrence of cancer, may be obtained from records, such as hospital records, cancer registrations, death certificates or some other specialized surveillance method (see Section 2.9). When records are used, the data available are limited to outcomes that are recorded routinely, their completeness, and the way in which they are coded.

Because malignancies develop slowly and are relatively rare, studies of the relationship between suspected carcinogenic exposures and cancer may require the observation of many participants over a long period. One way to avoid this is to use intermediate end-points as cancer surrogates: that is, to use as an outcome a biological event that is believed to lie on the causal pathway between exposure and cancer. Studies that use intermediate end-points are, in principle, quicker, smaller and less expensive than those using malignancy as the outcome. For instance, a study of the relationship between diet and estrogen metabolism could be carried out on several dozen patients, whereas a dietary intervention study with breast cancer as the end-point would require tens of thousands of women with many years of follow-up (Schatzkin *et al.*, 1990). The underlying assumption in these studies is that the observed relationship between exposure (e.g., diet) and the intermediate end-point (e.g., estrogen metabolism) reflects a similar relationship between exposure and the cancer of interest. Clearly, this assumption must be validated before the intermediate end-point can be used as a cancer surrogate (Toniolo *et al.*, 1997).

2.6 Validity and reliability of measurements of exposure and outcome

2.6.1 Validity

Validity is defined as the extent to which an instrument (for example, a questionnaire or a laboratory test) measures what it is intended to measure. Validity can be determined only if there is a reference procedure or 'gold standard': a definitive procedure to determine the characteristic being measured. For example, information on birth weight obtained from an interview can be validated against hospital records, and food-frequency questionnaires against food diaries and biological measurements. However, in some circumstances there is no obvious reference procedure and the best available method must be taken as the standard.

Consider the simple example of a test that can give only a positive or negative (i.e., binary) result. When the same subjects have been examined by both the study test and the gold standard, the findings can be expressed in a 2×2 table, as in Table 2.1.

The *sensitivity* of the study test is the proportion of individuals classified as positives by the gold standard who are correctly identified by the study test:

Sensitivity = $a/(a+c)$

		Gold standard	
		Positive	Negative
Study test	Positive	*a*	*b*
	Negative	*c*	*d*

a, true positives; *b*, false positives; *c*, false negatives; *d*, true negatives .

Table 2.1.
General layout of a 2 x 2 table to assess the validity of a test that can give only a binary result.

The *specificity* of the study test is the proportion of individuals classified as negatives by the gold standard who are correctly identified by the study test:

Specificity = $d/(b+d)$

The *predictive value of a positive study test result* represents the probability that someone with a positive study test result really has the characteristic of interest as determined by the gold standard:

Predictive value of a positive study test result = $a/(a+b)$

The *predictive value of a negative study test result* represents the probability that someone with a negative study test result does not have the characteristic of interest as determined by the gold standard:

Predictive value of a negative study test result = $d/(c+d)$

Example 2.1. *A variety of laboratory methods have been developed for detecting human papillomavirus (HPV) infection of the cervix uteri. In a study conducted some years ago, the performance of a new commercially available dot-filter hybridization test (ViraPap®) was assessed by comparing its results with those obtained using a gold standard test in a sample of 450 women who attended a clinic for sexually transmitted diseases in Washington state, USA during 1987–88 (Kiviat* et al.*, 1990). The Southern hybridization test, which is expensive and time-consuming, was taken as the gold standard in this study. The results are shown in Table 2.2.*

Table 2.2.
Comparison of ViraPap® and Southern hybridization methods in the diagnosis of cervical HPV infection in a sample of women who attended a sexually transmitted disease clinic.[a]

		Southern hybridization (gold standard test)		
		Positive	**Negative**	**Total**
ViraPap® (new test)	**Positive**	62	22	84
	Negative	7	359	366
	Total	69	381	450

[a] Modified from Kiviat *et al.*, 1990

These data yield the following for the ViraPap® test:
***Sensitivity** = 62/69 = 90%*
***Specificity** = 359/381 = 94%*
***Predictive value of a positive ViraPap® test** = 62/84 = 74%*
***Predictive value of a negative ViraPap® test** = 359/366 = 98%*

An ideal test has high sensitivity (correctly identifies a high proportion of truly exposed or diseased individuals) and high specificity (gives few positive results in unexposed or non-diseased individuals). In Example 2.1, the ViraPap® test had both high sensitivity and high specificity, indicating that the test was highly valid in the detection of cervical HPV infection (as compared to the Southern hybridization test) and therefore that its results would be little affected by measurement error.

While the predictive value of a study test result strongly depends upon the frequency of the disease (or other characteristic of interest) in the population, sensitivity and specificity are essentially unaffected. When the disease frequency changes, the numbers of diseased people as determined by the gold standard (left-hand column) change in proportion to the numbers of non-diseased people (right-hand column). Unlike sensitivity and specificity, the predictive value of a study test result depends on the numbers in both columns, and will change if the frequency of the disease changes.

Example 2.2. *Suppose that the same ViraPap® test was used in a sample of 450 apparently healthy women who visited their general practitioners for a regular check-up. The results are given in Table 2.3.*

		Southern hybridization (gold standard test)		
		Positive	**Negative**	**Total**
ViraPap® (new test)	**Positive**	21	26	47
	Negative	2	401	403
	Total	23	427	450

Table 2.3.
Comparison of ViraPap® and Southern hybridization methods in the detection of cervical HPV infection among apparently healthy women: hypothetical data.

These data yield the following for the ViraPap® test:
Sensitivity *= 21/23 = 91%*
Specificity *= 401/427 = 94%*
Predictive value of a positive ViraPap® test *= 21/47 = 45%*
Predictive value of a negative ViraPap® test *= 401/403 = 100%*

In Example 2.2, the predictive value of a positive ViraPap® test is markedly decreased (from 74% to 45%). This is because the proportion of HPV-infected women (as determined by the gold standard) was much higher (69/450 = 15%) in the sample of women who attended the clinic for sexually transmitted disease (Table 2.2) than among the group of apparently healthy women (23/450 = 5%). Thus, diagnostic tests which are useful in clinical medicine may perform poorly in epidemiological surveys or in population screening programmes. In clinical medicine, diagnostic tests are applied to patients in populations already selected as having a high occurrence of the condition. In this situation, the test may have

high predictive value. In an epidemiological survey of an unselected population, the same test may have poor predictive value because the frequency of the condition is much lower. For example, mammography has high predictive value as a test for breast cancer in women who consult doctors because of a lump in the breast, but low predictive value when used to screen apparently healthy women in the population. These issues are discussed further in Chapter 16.

The selection of a gold standard is a crucial aspect of evaluating the validity of any measurement. Unfortunately, in many cases there is no appropriate gold standard, and the investigator has to rely on the best available method. For instance, for many years, Southern hybridization was regarded as the gold standard method for detecting cervical HPV infection. However, with the development in recent years of polymerase chain reaction (PCR) to amplify HPV-specific DNA sequences, these newer methods have become the accepted gold standard.

Example 2.3. *The performance of the ViraPap® test was compared with that of the polymerase chain reaction (PCR) in newly diagnosed cervical cancer patients. Results are shown in Table 2.4.*

Table 2.4.
Comparison of ViraPap® and polymerase chain reaction (PCR) in the detection of cervical HPV infection.[a]

		PCR (gold standard test)		
		Positive	**Negative**	**Total**
ViraPap® test	**Positive**	163	11	174
	Negative	120	79	199
	Total	283	90	373

[a] From Muñoz *et al.* (unpublished)

These data yield the following for the ViraPap® test:
Sensitivity *= 163/283 = 58%*
Specificity *= 79 / 90 = 88%*

In Example 2.3 the validity of the ViraPap® test (as measured by its sensitivity and specificity) was much lower than when Southern hybridization was used as the gold standard method (Example 2.1). This is because the PCR method is more sensitive and more specific than the Southern hybridization technique.

Not all tests give a simple yes/no result. Some yield results that are numerical values along a continuous scale of measurement. In these situations, high sensitivity is obtained at the cost of low specificity and vice versa. For example, the higher the blood pressure, the more probable is hypertensive disease. If a diagnostic or screening test for hypertension is

set at a diastolic pressure of 90 mmHg, most hypertensive patients would be detected (high sensitivity) but many non-diseased subjects (with diastolic blood pressure higher than 90 mmHg) will be wrongly classified as hypertensive (low specificity). If the screening level for hypertensive disease is set at 110 mmHg for diastolic blood pressure, most non-diseased individuals would be excluded (high specificity), but many hypertensive patients (with diastolic blood pressures lower than 110 mmHg) would be missed (low sensitivity).

Example 2.4. *A new laboratory assay measuring the concentration of a particular enzyme in the blood is developed. To assess its value in the diagnosis of a specific cancer, the new test is applied to 360 hospital patients and the results are compared with those from anatomo-pathological examination. Blood concentrations of the enzyme ≥40 IU are taken as positive results. The results are shown in Table 2.5.*

		Anatomo-pathological examination (gold standard test)		
		Positive	**Negative**	**Total**
Blood assay	**Positive (≥40 IU)**	190	80	270
	Negative (<40 IU)	0	90	90
	Total	190	170	360

Table 2.5.
Comparison of a new laboratory assay with anatomo-pathological examination in the diagnosis of a specific cancer: hypothetical data.

The following can be calculated for the new laboratory assay:
Sensitivity *= 190/190 = 100%*
Specificity *= 90 / 170 = 53%*

In Example 2.4, other blood concentration values could be taken as cut-off values to define the assay results as 'positive' or 'negative'. Table 2.6 gives the sensitivity and specificity of the blood assay for different cut-off values. The sensitivity of the laboratory assay decreases as the cut-off value increases, whereas the reverse is true for specificity. This is clearly illustrated in Figure 2.1.

One way to summarize the validity of a continuous measurement is to plot sensitivity against (1 – specificity) for different cut-off values. This curve is called the receiver operating characteristic (ROC) curve. The ROC curve corresponding to the data in Table 2.6 is shown in Figure 2.2.

The closer the ROC curve of a particular test is to the top left-hand corner of the box, where both the sensitivity and specificity are maximized, the better the test. A test with a curve that lies on the diagonal is for practical purposes useless, and no better than a complete guess.

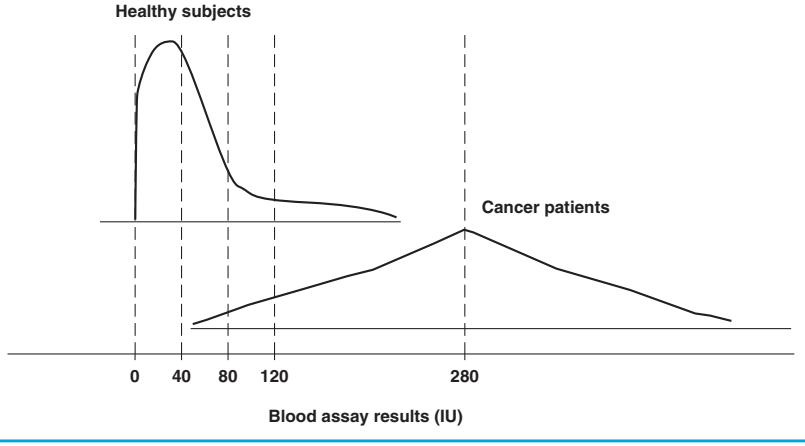
Table 2.6.
Sensitivity and specificity of the blood assay for different cut-off values: hypothetical data.

Cut-off value (IU)	Result of blood assay	Result of anatomo-pathological examination	Number of patients
40[a]	+	+	190
	+	–	80
	–	+	0
	–	–	90
80[a]	+	+	188
	+	–	42
	–	+	2
	–	–	128
120[a]	+	+	173
	+	–	25
	–	+	17
	–	–	145
280[a]	+	+	95
	+	–	0
	–	+	95
	–	–	170

[a] Blood assay results equal to or greater than the cut-off value were taken as positive: +, positive result; –, negative result

40 IU:	sensitivity = 190 / 190 = 100%	specificity = 90 / 170 = 53%
80 IU:	sensitivity = 188 / 190 = 99%	specificity = 128 / 170 = 75%
120 IU:	sensitivity = 173 / 190 = 91%	specificity = 145 / 170 = 85%
280 IU:	sensitivity = 95 / 190 = 50%	specificity = 170 / 170 = 100%

Figure 2.1.
The upper curve describes the distribution of results of the blood assay among healthy individuals and the lower curve the distribution among cancer patients (as defined by the anatomo-pathological examination). Different cut-off values are used to classify the results of the blood assay as 'positive' or 'negative'.



2.6.2 Reliability

Reliability, sometimes also called repeatability or reproducibility, is a measure of the consistency of the performance of a test when used under similar circumstances. To be valid, a measurement must be reliable. However, reliability is not in itself sufficient for validity: in other words, a test may yield the same result consistently, but the result may not be the true (valid) one. Poor reliability of a measurement may be due to variation when a subject is tested on different occasions (biological variation), or to errors in the measurement technique (observer and instrument variation). Checks of the repeatability of measurements of the main exposures and outcomes should usually be included in an epidemiological study. These checks can take various forms.

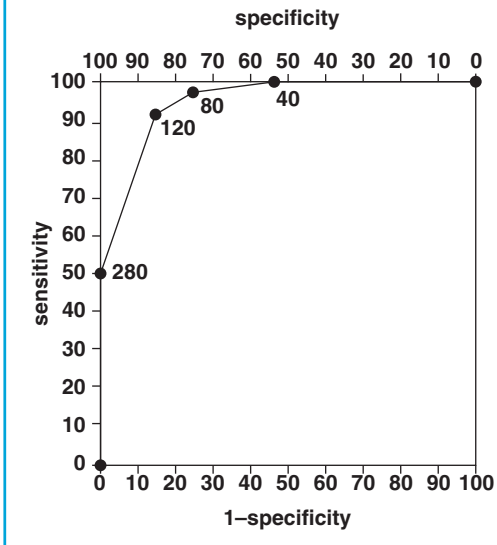


Figure 2.2.
Receiver operating characteristic (ROC) curve for the data in Table 2.6 and Figure 2.1.

(1) Intra-observer or intra-measurement reliability

Intra-observer or intra-measurement reliability can be determined by having the same observer perform the same measurements on the same subjects on two or more separate occasions. For example, data from medical records may be extracted by the same abstractor on two occasions; the same interviewer may re-interview subjects after a time interval; duplicate biological samples may be re-processed by the same laboratory technician. These separate measurements are then compared. The appropriate time interval between measurements varies according to the type of outcome or exposure measurement. If it is too short, subjects and/or observers may recall the previous result; if it is too long, the subject's exposure or outcome status may have changed (of course, this is not a problem when data are extracted from medical records).

(2) Inter-observer reliability

Inter-observer reliability can be assessed by having the same subjects measured by two or more independent observers. For example, the performance of two or more data abstractors may be compared using information extracted independently from the same medical records, or the performance of two or more interviewers may be compared using independent interviews of the same subjects on two different occasions. Again, the interval used between measurements needs careful consideration.

Consider the simple example of a test that can give only a positive or negative (i.e. binary) result. The agreement between pairs of measurements carried out by two independent observers on the same subjects can be presented as a 2×2 table (Table 2.7).

One measure of repeatability is the observed agreement (O) or mean pair agreement index, which can be calculated as:

(No. of agreements/Total no. of pairs) = $(a + d) / N$

Table 2.7.
General layout of a table to assess reliability between two observers for a binary test. *a*, *b*, *c* and *d* refer to the numbers of pairs of observations where the observers gave the indicated result.

		Observer B		
		Positive	Negative	Row total
Observer A	Positive	*a*	*b*	*a*+*b*
	Negative	*c*	*d*	*c*+*d*
Column total		*a*+*c*	*b*+*d*	
			Grand total	*N*

This index has the disadvantage that some agreement would be expected even if both observers simply guessed the result. The kappa statistic (κ) is an alternative measure that takes account of the agreement expected solely on the basis of chance.

To calculate the kappa statistic, the number of pairs of observations that would be expected on the basis of chance in cells (++) and (– –) must first be calculated. The expected value in any cell is given by:

[(Total of relevant row) × (Total of relevant column)]/Grand total

Thus for cell (++), the expected value will equal:

$[(a+b) \times (a+c)]/N$

and for cell (– –), the expected value will equal:

$[(c+d) \times (b+d)]/N$

The expected agreement on the basis of chance (E) can now be calculated as:

[Expected value for cell (++) + Expected value for cell (– –)] / N

The actual agreement beyond chance is therefore:

Observed agreement (O) – Expected agreement (E)

This value is, however, difficult to interpret, as similar results may be obtained for different values of O and E. For instance, the actual agreement beyond chance is equal to 0.20 for values of $O = 0.95$ and $E = 0.75$, and for $O = 0.75$ and $E = 0.55$. What we need to know is how much does it represent in relation to the maximum potential agreement beyond chance that could have been achieved. Complete agreement would imply that all the results would have fallen in cells (++) and (– –) and, therefore, $(a+d)/N$ would have equalled 1. Thus, the potential for agreement beyond chance is

$1 - E$

The kappa statistic indicates how much the actual agreement beyond chance (O–E) represents relative to this potential ($1 - E$).

$$\text{Kappa } (\kappa) = (O - E)/(1 - E)$$

The kappa statistic can be used in a similar way to measure intra-observer variability. The values of this coefficient may vary from –1.0 to 1.0. A value of 1.0 indicates perfect agreement and a value of zero means agreement is no better than would be expected on the basis of chance alone; a negative value indicates that the level of disagreement is greater than that expected on the basis of chance. While there is no value of kappa that can be regarded universally as indicating good agreement, in practice, a κ value considerably less than 0.5 indicates poor agreement. Landis and Koch (1977) suggested the following guidelines: kappa values ≤0.40 represent poor-to-fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.

Use of the kappa statistic can be extended to situations where the results of the test are classified in more than two categories, as in Example 2.5. The kappa shows substantial agreement between observers A and B. Intra-observer agreement was calculated in a similar way: the kappa statistic equalled 0.83. In general, intra-observer agreement tends to be better than inter-observer agreement.

kappa values should not be presented alone, as they provide a summary measure of agreement without giving any indication where disagreements occurred. The results of a reliability study should therefore always be presented in a table similar to Table 2.8, so that the main areas of agreement and disagreement are apparent. If different importance is given to different types of agreement or disagreement, the kappa statistics may be weighted to take this into account (Landis & Koch, 1977).

Methods are also available for assessing the reliability of measurements that provide results on a continuous scale (e.g. blood pressure measurements, blood glucose levels): however, these are beyond the scope of this chapter. A discussion of these methods can be found in Bland & Altman (1986).

2.7 Consequences of measurement error

Errors in measurement can lead to individuals being misclassified and to spurious conclusions about the relationship between the exposure and the outcome. The impact of measurement errors on the results of an epidemiological study depends essentially on the nature of any misclassification.

Consider the following example. Suppose that to determine whether cigarette smoking is associated with lung cancer, we rely on a questionnaire that asks ‘Have you ever smoked?’ and ‘Do you have lung cancer?’. The questionnaire is administered to 10 000 men. Assume that the ‘true’

Example 2.5. *In the study by Kiviat* et al. *(1990) mentioned in Section 2.6.1, the authors state: 'To assess inter-observer variation all autoradiographs were initially reviewed independently by two observers without their knowledge of other laboratory and clinical data, and specimens were classified as positive, negative, or indeterminate according to the manufacturer's specifications. Intra-observer variation was assessed by having membranes re-read by observer A six months later without her knowledge of other (or previous) laboratory or clinical data'. The results for inter-observer variability are given in Table 2.8.*

Table 2.8.
Inter-observer variability in the reading of the ViraPap® test.[a]

	Observer B[b]			
	Positive	**Negative**	**Indeterminate**	**Row total**
Observer A[b]				
Positive	58 (a_1)	8 (a_2)	2 (a_3)	68 (a)
Negative	12 (b_1)	357 (b_2)	3 (b_3)	372 (b)
Indeterminate	0 (c_1)	0 (c_2)	7 (c_3)	7 (c)
Column total	70 (n_1)	365 (n_2)	12 (n_3)	447 (N)

[a] Data from Kiviat *et al.* (1990).
[b] Figures represent numbers of pairs of observations where the observers gave the indicated result; letters in parentheses indicate each specific cell in the table.

Observed agreement $(O) = (a_1 + b_2 + c_3)/N = (58 + 357 + 7)/447 = 0.94$
Expected value for cell $a_1 = (a \times n_1)/N = (68 \times 70)/447 = 10.65$
Expected value for cell $b_2 = (b \times n_2)/N = (372 \times 365)/447 = 303.76$
Expected value for cell $c_3 = (c \times n_3)/N = (7 \times 12)/447 = 0.19$
Agreement expected on the basis of chance $(E) = (10.65 + 303.76 + 0.19)/447 = 0.70$
Actual agreement beyond chance $(O - E) = 0.94 - 0.70 = 0.24.$
Potential agreement beyond chance $= 1 - 0.70 = 0.30.$
Kappa $(\kappa) = 0.24 / 0.30 = 0.80$

Table 2.9.
Distribution of a population by smoking and disease status as determined by a perfect test for measuring smoking habits (sensitivity = 100%; specificity = 100%): hypothetical data.

		Cigarette smoking		
		Ever	**Never**	**Total**
Lung cancer	**Yes**	150	50	200
	No	1850	7950	9800
	Total	2000	8000	10 000

smoking status in this study population (as determined by a perfect test, having both a sensitivity and a specificity of 100%) is as indicated in Table 2.9. This table shows that lung cancer is more common among people who have smoked (ever smokers) (150 of 2000 = 7.5%) than among those who have never smoked (never smokers) (50 of 8000 = 0.63%). Thus, if a perfect method could be used to measure smoking habits in this example, ever smokers would be found to be 12 times (7.5% / 0.63% = 12) more likely to develop lung cancer than never smokers.

		Cigarette smoking		
		Ever	Never	Total
Lung cancer	Yes	150 – 0.2 × 150 = 120	50 + 0.2 × 150 = 80	200
	No	1850 – 0.2 × 1850 = 1480	7950 + 0.2 × 1850 = 8320	9800
	Total	1600	8400	10 000

Table 2.10. Distribution of a population by smoking and disease status as determined by a test for measuring smoking habits that has a sensitivity of 80% and a specificity of 100%: hypothetical data.

Suppose now that when the questionnaire is applied, 20% of smokers, regardless of their disease status, answered that they had never smoked (sensitivity=80%), but that all men who have never smoked reported this accurately (specificity=100%). The results that would be obtained with this imperfect questionnaire are shown in Table 2.10.

Using this imperfect questionnaire, the proportion of lung cancers in 'smokers' is 120/1600=7.5%. This is about eight times the proportion in 'never smokers' (80/8400=0.95%). Despite the poor quality of the data on smoking elicited by the questionnaire, the relationship between cigarette smoking and lung cancer, while appearing weaker than it truly is, is still evident.

Non-differential misclassification occurs when an exposure or outcome classification is incorrect for equal proportions of subjects in the groups being compared. In other words, the sensitivity and specificity of the exposure (or outcome) measurement are equal for both the diseased and non-diseased (or exposed and unexposed). In these circumstances, the misclassification is random (i.e., all individuals have the same probability of being misclassified).

In non-differential misclassification, individuals are wrongly classified, reducing the confidence that can be placed in each particular test result. Although this random misclassification has important implications in clinical medicine, it is of less concern in epidemiology, where groups rather than individuals are the main interest. Herein lies a great strength of epidemiology. In the above example, the association between smoking and lung cancer was weakened because those classifying themselves as 'never smokers' were in fact a mixture of those who had never smoked and those who had. Although this type of misclassification makes it more difficult to reveal an association between the exposure and the outcome of interest, the problem can usually be overcome by increasing the sample

size and/or replicating measurements (except, as discussed in Chapter 13, where there is non-differential misclassification of confounding variables). Thus, the epidemiologist can rely on simple, cheap and non-invasive tests which, despite being in general less valid than those used in clinical settings, are more appropriate for studies in the community.

This is an important aspect of epidemiological research that clinicians often find difficult to accept. Clinicians focus on individual patients, trying to obtain the most complete and valid information on which to base the most accurate diagnosis possible and the optimal treatment. Being accustomed to using specialized and high-technology procedures, they may find it hard to believe that one could undertake scientific studies based on relatively low-quality data such as those derived from questionnaires or death certificates.

Differential misclassification occurs when the sensitivity and/or specificity of the exposure measurement for the diseased group differs from that for the non-diseased group, or when the sensitivity and/or specificity of the outcome measurement for the exposed group differs from that for the unexposed group. In other words, differential misclassification may occur when errors in classification of outcome status are dependent upon exposure status, or vice versa. For example, clinicians may be more likely to diagnose leukaemia in children who live around nuclear power stations than in those living elsewhere, and women with breast cancer may be more likely to remember having taken oral contraceptives in the past than healthy women. In the example already considered, differential misclassification would have occurred if men with lung cancer were likely to report their smoking habits more or less accurately than men without lung cancer; in such circumstances, the resulting data could exaggerate, attenuate, or even reverse the relationship, and make the results misleading.

Differential misclassification is a consequence of defects in the design or execution of an epidemiological study. Unfortunately, it cannot be controlled for in the analysis, and its effect cannot be minimized by increasing the sample size.

A more detailed discussion of the consequences of errors in the measurement of exposure and outcome in the interpretation of epidemiological studies is given elsewhere in this book; in particular, in Chapter 13.

2.8 How can misclassification of exposure and outcome be reduced?

All procedures used in the measurements should be described in sufficient detail in the study protocol to allow reproduction of the measurements, within the limits of biological and physical variability, by other investigators. The protocol should include not only a description of the method of measurement, but also instructions for its application. All other procedures involved should also be specified.

For a personal interview, this will include:
- specifications for the training of interviewers and instructions given to them,
- instructions or explanations given by interviewers to subjects,
- the questionnaire used to elicit data from the subjects,
- quality-control procedures.

For a laboratory test, this will include:
- procedures for the preparation of subjects,
- procedures for the collection, manipulation, transport, and storage of the specimens,
- analytical procedures in the laboratory,
- quality-control procedures.

The epidemiologist should establish and maintain close contact with the specialists in the laboratory, so that standard criteria for collecting, storing and analysing specimens are established at the beginning of the study. Although most laboratories routinely apply intra- and inter-laboratory quality-control procedures, epidemiologists should send specimens without revealing the exposure (or disease) status of the subjects from whom they were collected. It is also advisable to send replicate samples without the laboratory staff being aware that they are replicates.

Measurement procedures should always be evaluated in a pilot study to identify any potential problems, gauge their validity and reliability, and determine in what way observers or responders may be biased. These issues are discussed further elsewhere in this book; in particular, in Chapters 13 and 18.

2.9 Sources of routine data

'Routine data' are derived from established data collection systems associated with the health and social services. In general, the data are not collected with the aim of answering any specific question. For whatever purpose they were collected, such data can often be used in epidemiological studies; these include data from censuses and population registers, birth and death certificates, cancer registrations, health information systems, medical and hospital records, etc. (see Chapter 11).

Routine data collection systems can provide information on the exposure(s) and outcome(s) of interest in an epidemiological study. Two such systems—death certification and cancer registration—are particularly important in cancer epidemiology.

2.9.1 Death certification

Mortality data are usually based upon a standard death certificate, which records the date of death, cause of death, age, sex, date of birth and place of residence of the deceased. In addition, occupation and other information may be recorded. In most countries, death certificates are usually completed by a doctor or other health worker but in some cases this is done by the police or other authorities. Once certificates are completed, the cause of

death is coded according to the *International Classification of Diseases*, now in its tenth revision (WHO, 1992). This is a hierarchical classification of diseases, from broad categories down to a detailed four-character classification (see Appendix 2.2). Usually, only the underlying cause of death is coded and used in mortality statistics, although contributing causes may also be coded.

While more complete and reliable than many routine sources of morbidity data, mortality data are still subject to some misclassification (Cameron & McGoogan, 1981; Heasman & Lipworth, 1966). A large international study of 8737 cancer deaths in cities in England, USA and Latin America revealed that of deaths classified on the death certificate as caused by cancer, 20% were due to other causes (Puffer & Griffith, 1967). However, 24.6% of cancer deaths had been wrongly classified under other causes of death. On balance, therefore, total cancer mortality was only 4% underestimated in the official statistics derived from death certificates. The degree of misclassification varied with cancer site, being greater for those that are more difficult to diagnose, such as primary liver cancer and brain tumours.

International cancer mortality statistics are published regularly by the World Health Organization (*World Health Statistics Annual* series) and by Segi and his colleagues (Kurihara *et al*., 1989).

2.9.2 Cancer registration

There are two types of cancer registry: hospital-based and population-based. Hospital-based cancer registries record all cancer patients seen in a particular hospital. Their main purpose is to contribute to patient care and administrative management, although they may be useful to a certain extent for epidemiological purposes. For instance, 'rolling' case–control studies may be set up to investigate the etiology of a particular cancer; this is achieved by comparing the characteristics of such cases with those of a control group, which may be made up of patients either with other types of cancer, or with other illnesses. Nevertheless, hospital-based registries cannot provide measures of the occurrence of cancer in the general population, because it is not possible to define the population from which cases arise.

Population-based cancer registries seek to record all new (incident) cancer cases that occur in a well defined population. As a result, they provide measures of the occurrence of cancer in their catchment population. Population-based cancer registration has been developed in many countries to provide reasonably comparable data on cancer incidence and as a resource for epidemiologal studies. Cancer incidence data from higher-quality registers are compiled by the International Agency for Research on Cancer in the series *Cancer Incidence in Five Continents* (Doll *et al.*, 1966; Waterhouse *et al.*, 1970, 1976, 1982; Muir *et al.*, 1987; Parkin *et al.*, 1992, 1997). Some indicators of data quality for the different registries included in this publication are tabulated in these volumes. However, these are mostly indirect indicators of data quality: proportion of registrations ver-

ified histologically; proportion of cases registered on the basis of information on the death certificate only; proportion of cases with missing information, etc. More systematic analyses of the validity of cancer registration data are available for certain registries, where a sample of cases was re-abstracted and re-coded (see Parkin *et al.*, 1994).

The majority of cancer registries collect information about cancer patients, such as their occupation, social class, country of birth, ethnicity, etc. Occurrence of cancer can therefore be examined in relation to these variables.

The role of cancer registries in cancer epidemiology is discussed in detail in Chapter 17.

2.9.3 Record linkage

Information on individuals from birth to death is available in the records of many institutions and agencies. These various records may be merged into a single comprehensive record using personal identifiers, in a process known as record linkage. The unified record can then be used in epidemiological and public health investigations. The potential for linkage between registers varies enormously between countries according to how the relevant information is collected and identified. Thus, in the Nordic countries, where everyone is assigned a personal number which is used for all social security, census and health records, mortality and cancer incidence data can readily be traced and linked to other data-sets of interest. In the United Kingdom, a national register linked to the health service is widely used for follow-up studies of cancer and mortality, and computerized linkage is now possible for people who were alive in January 1991, matching information such as name and date of birth.

Linkage of cancer registry records with records from other sources, such as census data and company records, has been undertaken in an attempt to investigate risk factors for occupational cancers and cancers of the reproductive system. Registries can also draw information on exposure from hospital records, as they often record hospital admission numbers. This linkage with hospital records has been used in studies of cancer risks associated with radiotherapy and other treatments (Day & Boice, 1983; Kaldor *et al.*, 1992).

The Oxford Record Linkage Study (ORLS) and the national Scottish medical record linkage system are two good examples of record linkage. The ORLS was established in Oxford, UK in 1962, to assess the feasibility, cost and methods of medical record linkage for an entire community. The system links morbidity and mortality data and provides information on a wide range of variables. Data in the system can be used to study etiological questions and to assess the natural history of various diseases (Acheson, 1967; Baldwin *et al.*, 1987). In Scotland, all births, deaths, hospitalizations, cancer incidence, school medical examinations and handicapped children's records can be linked (Heasman & Clarke, 1979). Similar record-linkage systems have been set up in the USA by the National Center for Health Statistics (Feinleib, 1984) and in many other developed countries.

Further reading

* Comprehensive coverage of principles and practical aspects of questionnaire design, the conduct of personal interviews, the abstraction of information from records, and the use of biological measurements and measurements in the environment, is given by Armstrong *et al.* (1992). Although this book focuses on exposure measurement, many of the principles presented are also relevant to the measurement of outcomes.

* An often-referenced paper on the validity and reliability of tests that yield results on a continuous scale (e.g., blood pressure measurements) is that by Bland & Altman (1986).

* For a further, more complex, discussion on the kappa statistic, see Feinstein & Cicchetti (1990), Cicchetti & Feinstein (1990) and Lantz & Nebenzahl (1996).

Box 2.1. Key issues

- In epidemiological studies, it is necessary to measure: (1) the primary exposure(s) of interest; (2) other exposure(s) that may influence the outcome (potential confounders); and (3) the outcome(s) of interest.

- Many approaches can be used to measure exposure and outcome. These include personal interviews, self-administered questionnaires, diaries, records, biological measurements and measurements in the environment. Each method has its own advantages and disadvantages.

- In any epidemiological study, it is important to assess the validity and reliability of the main measurements of exposure and outcome. This will provide an estimate of the magnitude of measurement errors and their probable impact on the study results. Measurement errors may be non-differential or differential.

- *Non-differential measurement error* occurs when the sensitivity and specificity of the exposure measurement for the diseased group equal those for the non-diseased group, or when the sensitivity and specificity of the outcome measurement is the same for both exposed and unexposed subjects. Non-differential measurement error generally leads to under-estimation of the association between the exposure and the outcome. Although non-differential measurement errors make it more difficult to reveal an association between the exposure and the outcome, this can usually be overcome by increasing the sample size and/or replicating measurements.

- *Differential measurement error* occurs when the sensitivity and/or specificity of the exposure measurement for the diseased subjects differs from that for the non-diseased subjects, or when the sensitivity and/or specificity of the outcome measurement is different for exposed and unexposed subjects. Differential measurement error can exaggerate, attenuate, or even reverse, the relationship between the exposure and the outcome, so that the results of the study can be misleading. Unfortunately, differential measurement errors cannot be controlled for in the analysis, and their effects cannot be lessened by increasing the sample size.

Appendix 2.1
Designing a questionnaire

Questionnaires are used in epidemiology to assess exposure levels to possible causal agents and, less often, to determine the presence or absence of disease, or another outcome of interest.

A2.1.1 Objectives of questionnaire design

(1) To provide valid measurements of the exposure(s) and outcome(s) being studied.
(2) To design a questionnaire that is easily completed by the interviewer and/or subject.
(3) To facilitate data-processing and analysis.

A2.1.2 General principles of questionnaire design

Content

The questionnaire should be as brief as possible, with every question being carefully justified in terms of the objectives of the study. It is important to ensure that the variables needed for the analysis can be easily obtained from the questionnaire.

Types of question

There are two main types of question: 'open-ended' and 'closed-ended'. *Open-ended questions* allow the respondents to answer on their own terms and should be recorded in the respondent's own words. Open-ended questions should be used for numerical data (for example, age, date of birth) and for questions having many possible answers (e.g. country of birth).

Example A2.1.1. *Example of an open-ended question.*

What is your mother tongue?..... ☐

Closed-ended questions allow only a limited range of answers. The questionnaire should specify in detail all the possible alternative answers. With multiple alternative answers, a final alternative 'Other: please specify...' should be provided unless it is certain that all possible answers have been provided. A 'Do not know' option should also be given for questions where it is possible that some subjects may not know (or may not remember) the answer. A 'Not applicable' option should be given if the question does not apply to all subjects.

Example A2.1.2. *Example of a closed-ended question.*

7. If you have NEVER BEEN PREGNANT was it because:

You never tried	☐ 1
You tried but it never happened	☐ 2
Other reasons: please specify	☐
Not applicable	☐ 7

Epidemiological questionnaires usually contain a majority of closed-ended questions to reduce the possibility of interviewer, response, interpretation, and/or coding bias, and to facilitate data-processing.

Wording of questions

Questions must be written in simple, non-threatening language, avoiding the use of abbreviations and technical jargon. The wording should avoid any suggestion that a particular answer is preferred by the researcher(s). Each question should contain only one concept related to a clear time period.

Order of questions

Questions should follow a logical sequence resembling, as far as possible, the sequence that the respondents might expect to follow when thinking about the topic. Questions about a particular subject should be grouped together, and proceed from the general to the particular. When a response to a general question makes further responses on that topic irrelevant (e.g., a woman who has never been pregnant need not answer questions about number and characteristics of pregnancies), a branching of the question sequence may be introduced. This should be as simple as possible, with clear instructions given on the questionnaire (Example A2.1.3).

Questionnaire layout

Layout is important in both self- and interviewer-administered questionnaires. A pleasant appearance will arouse interest and encourage correct completion. A separate page with a brief introduction, explanatory notes and instructions should precede the first question. To help interviewers and subjects, long questionnaires may be subdivided into sections, each one corresponding to a specific topic. All questions should be assigned a number.

If some questions are optional, this should be indicated on the questionnaire with clear instructions and appropriate branch and jump expla-

Example A2.1.3. *Example of instructions for omitting questions (jumping).*

1. Have you ever been pregnant?

Yes ☐ 1

No ☐ 2

If No, please go to question 4.

If Yes,

2. How many pregnancies in total (including still births, miscarriages and abortions) have you had? ☐☐

nations. For questions that are repeated several times, such as questions about each pregnancy, a tabular layout may be used (Example A2.1.4).

Space should be provided at the end of the questionnaire for any information or comments that the subject may wish to add.

Example A2.1.4. *Example of a question with a tabular layout.*

3. Please indicate the characteristics of your pregnancies

	1st	2nd	3rd	4th	5th
Age at start of pregnancy (years)	☐☐	☐☐	☐☐	☐☐	☐☐
Outcome	Birth ☐ 1 Still birth ☐ 2 Miscarriage ☐ 3 Abortion ☐ 4	Birth ☐ 1 Still birth ☐ 2 Miscarriage ☐ 3 Abortion ☐ 4	Birth ☐ 1 Still birth ☐ 2 Miscarriage ☐ 3 Abortion ☐ 4	Birth ☐ 1 Still birth ☐ 2 Miscarriage ☐ 3 Abortion ☐ 4	Birth ☐ 1 Still birth ☐ 2 Miscarriage ☐ 3 Abortion ☐ 4
Duration of pregnancy (weeks)	☐☐	☐☐	☐☐	☐☐	☐☐
Breast-fed	Yes ☐ 1 No ☐ 2 Not applicable ☐ 7 Do not know ☐ 9	Yes ☐ 1 No ☐ 2 Not applicable ☐ 7 Do not know ☐ 9	Yes ☐ 1 No ☐ 2 Not applicable ☐ 7 Do not know ☐ 9	Yes ☐ 1 No ☐ 2 Not applicable ☐ 7 Do not know ☐ 9	Yes ☐ 1 No ☐ 2 Not applicable ☐ 7 Do not know ☐ 9

Method of administration

The questionnaire can be either self-administered or interviewer-administered. In general, self-administered questionnaires must be simpler and much more carefully designed than those intended for use by interviewers.

Recording and coding of responses

Most questionnaires will be prepared to allow numerical coding of all responses for processing by computer. Every possible answer on the form is assigned a code (as in Examples A2.1.2 to A2.1.4). Numerical data (e.g., number of pregnancies) do not require coding, as the exact number can be entered. But even with such pre-coded questionnaires, some coding of data collection will still be required for some open-ended questions or for the 'Other: please specify' category of closed-ended questions (as in Example A2.1.2).

Coding of questionnaires may be a complex task and it may be necessary to develop a coding manual with specific coding rules. Various classification systems have been developed and published which can be used to code cancers by their topography (e.g., *International Classification of Diseases,* WHO, 1992) and by their morphology and behaviour (e.g., *International Classification of Diseases for Oncology*, Percy *et al.*, 1990) (see Appendix 2.2), occupations (e.g., *Classification of Occupations*, OPCS, 1970), and many other variables.

A2.1.3 Evaluation of a questionnaire

Questionnaires should be subject to two forms of evaluation: pre-testing and assessment of validity.

Pre-testing

All questionnaires should be pre-tested. This involves testing the draft questionnaire on samples of subjects similar to those who will ultimately be studied. Its purpose is to identify questions that are poorly understood, ambiguous, or evoke hostile or other undesirable responses. Pre-tests should be carried out using the same procedures that will finally be used in administering the questionnaire. Interviewers and study subjects should be asked to provide feedback and the questions revised in the light of their comments. Several rounds of pre-testing will usually be necessary before the final form of a questionnaire is developed.

Assessment of validity

The validity of the questionnaire as a measure of the variables of interest should always be determined in a sample of subjects before the main study is undertaken. This requires comparison of the results obtained using the questionnaire with those obtained using a gold standard test (see Section 2.6). For instance, questions on past hospitalizations and surgical interventions may be validated against hospital records. Validation is usually difficult, often expensive, and may sometimes be impossible, when no appropriate gold standard is available.

A2.1.4 Use of standard questionnaires

If a standard questionnaire for measurement of a particular exposure is available, it may be best to use this, rather than spending time and effort designing a new one. Moreover, the standard questionnaire will have been used extensively and proved satisfactory, and may even have been validated (although validity in one population may not ensure validity in another). Use of a standard questionnaire will also allow comparison of the data gathered with those collected in other studies.

Some changes in the format of a standard questionnaire may be needed to make it suitable for a particular study population. Be aware that such changes may affect the validity of the questionnaire; however, any modification can be tested for validity against the original questionnaire.

A full discussion of questionnaire design is given by Armstrong *et al.* (1992).

Appendix 2.2
Classification of diseases

Neoplasms can be classified in many ways, but the most important classifications for the epidemiologist are those based on:

(1) Topography—the site in the body where the tumour is located.
(2) Morphology (or histology)—the microscopic characteristics of the tumour.
(3) Behaviour—the tendency to invade other tissues (malignant, benign, *in situ,* and uncertain).

Uniform definitions and uniform systems of classification are fundamental to the quantitative study of diseases. Without a standard classification tool that remains fixed for periods of time and is applied uniformly, meaningful comparative analyses of morbidity and mortality data would be impossible. The *International Classification of Diseases* (ICD), published by the World Health Organization, is such a standard classification tool. It is revised every ten years or so (Table A2.2.1); the 10th revision (ICD-10) (WHO, 1992) is currently in use. An historical review of disease classification from the first revision, the Bertillon Classification of Causes of Death, until 1947 can be found in the introduction to ICD-7 (WHO, 1957), and an account of classification in the years 1948–1985 is given by Muir and Percy (1991).

Revision	Publication year	Publisher
1st (ICD-1)	1900	French Government
2nd (ICD-2)	1910	
3rd (ICD-3)	1920	
4th (ICD-4)	1929	Health Organization of the League of Nations
5th (ICD-5)	1938	
6th (ICD-6)	1948	World Health Organization
7th (ICD-7)	1957	
8th (ICD-8)	1967	
9th (ICD-9)	1977	
10th (ICD-10)	1992	

Table A2.2.1.
Revisions of the International Classification of Diseases.

Although retaining the traditional structure of ICD-9, the 10th revision of the ICD uses an alphanumeric coding scheme—the first character of the category is a letter—replacing the numeric codes of ICD-9 and previous revisions. This change provides a larger coding frame and leaves scope for future

inclusion of new disease entities without disrupting the numbering system.

ICD-10 has three volumes. Volume 1 deals with the tabular list of classification at the level of three and four characters, special tabulations of morbidity and mortality, and definitions and nomenclature regulations. Volume 2 is essentially an instruction manual. Volume 3 contains an alphabetical index.

The ICD chapter that deals with neoplasms presents a primarily topographic classification arranged according to the anatomical site of the tumour, with the exception of a few histological types such as lymphomas and leukaemias (Table A2.2.2). Organs are ordered according to organ systems. Neoplasms with a given behaviour are grouped as malignant, benign, *in situ* and of uncertain behaviour.

Table A2.2.2.
Classification of neoplasms according to ICD-10 (WHO, 1992).

C00-C75	Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue	
	C00-C14	Lip, oral cavity and pharynx
	C15-C26	Digestive organs
	C30-C39	Respiratory and intrathoracic organs
	C40-C41	Bone and articular cartilage
	C43-C44	Skin
	C45-C49	Mesothelium and soft tissue
	C50	Breast
	C51-C58	Female genital organs
	C60-C63	Male genital organs
	C64-C68	Urinary tract
	C69-C72	Eye, brain and other parts of central nervous system
	C73-C75	Thyroid and other endocrine glands
C76-C80	Malignant neoplasms of ill-defined, secondary and unspecified sites	
C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue	
C97	Malignant neoplasms of independent (primary) multiple sites	
D00-D09	*In situ* neoplasms	
D10-D36	Benign neoplasms	
D37-D48	Neoplasms of uncertain or unknown behaviour	

The first morphological classification was developed in 1951 and many others have since emerged (Table A2.2.3). The *Manual of Tumor Nomenclature and Coding* (MOTNAC) (American Cancer Society, 1951; Percy *et al.*, 1968) and, more recently, the *International Classification of Diseases for Oncology* (ICD-O) (WHO, 1976; Percy *et al.*, 1990) have been the most widely used. They provide not only morphology and behaviour codes, but also topography codes that are directly related to the ICD codes. A full discussion of the merits and drawbacks of each of these classifications is given by Muir and Percy (1991).

Publication year	Morphological code manual	Publisher	Main characteristics
1951	*Manual of Tumour Nomenclature and Coding (MOTNAC)* 1st edition	American Cancer Society	Morphology codes Behaviour codes
1956	*Statistical code for Human Tumours (STAT CODE)*	World Health Organization	Topography codes from ICD-7 Morphology codes from MOTNAC Behaviour codes from MOTNAC
1965	*Systematized Nomenclature of Pathology (SNOP) (Section 8,9 – neoplasms)*	College of American Pathologists	Topography codes unrelated to ICD Morphology codes
1968	*Manual of Tumor Nomenclature and Coding (MOTNAC)* 2nd edition	American Cancer Society	Topography codes from ICD-8 Morphology codes from SNOP
1976	*ICD-O,* 1st edition	World Health Organization	Topography codes from ICD-9 Morphology codes from MOTNAC (with one-digit extension) Behaviour codes from MOTNAC
1977	*Systematized Nomenclature of Medicine (SNOMED) (Section 8,9 – neoplasms)*	College of American Pathologists	Review of SNOP Topography codes unrelated to ICD Morphology codes from ICD-O
1990	*ICD-O,* 2nd edition	World Health Organization	Topography codes from ICD-10 Morphology codes from ICD-O, 1st edition Behaviour codes from ICD-O, 1st edition

Table A2.2.3.
Morphology and behaviour classifications of neoplasms.

The major advantage of ICD is that it is truly international, being used by all WHO Member States for tabulating the causes of death and for most health statistics. The main disadvantage is that, for the majority of sites, no separation on the basis of morphology is possible. As a result, it is generally recommended that agencies interested in identifying both the site and morphology of tumours, like cancer registries and pathology laboratories, use ICD-O, which is a dual-axis classification providing independent coding systems for topography and morphology.

As new classifications and new revisions of ICD and ICD-O have come into use, data coded by previous classifications must be converted to the new codes. The National Cancer Institute of the USA has produced a series of conversion tables for neoplasms (e.g., Percy, 1980, 1981, 1983; Percy & van Holten, 1979). Summary tables of equivalence between various revisions of the ICD have also been published in certain volumes of *Cancer*

Incidence in Five Continents (e.g., Waterhouse *et al.*, 1976; Muir *et al.*, 1987). Programs that perform conversions from ICD-O (1st edition) to ICD-O (2nd edition) and vice versa, from ICD-O (1st and 2nd editions) to ICD-9, and from ICD-O (2nd edition) to ICD-10 have been developed by the International Agency for Research on Cancer (IARC) and are available on diskette for use on microcomputers (Ferlay, 1994).