Chapter 16
Cancer prevention

16.1 Introduction

The preceding chapters of this book have focused on principles and methods needed to study the determinants of disease and their effects. The ultimate goal of epidemiology, however, is to provide knowledge that will help in the formulation of public health policies aimed at preventing the development of disease in healthy persons.

Nearly all important issues in cancer prevention are linked to the natural history of the disease 'cancer', which can be summarized as shown in Figure 16.1. Here, point A indicates the biological onset of the disease and the start of the pre-clinical phase. This may be the point at which an irreversible set of events (e.g., gene mutation) takes place. As a result of progression of the disease, symptoms and/or signs appear that bring the patient to medical attention and diagnosis at point C. This is the end of the pre-clinical phase, which is the period from A to C, and the beginning of the clinical phase of the natural history. The disease may then progress to cure (D_1), to permanent illness and disability (D_2) or to death (D_3). The time from initial symptoms and/or signs to cure, permanent illness or death may reflect the effects of treatments given, as well as the underlying characteristics of the untreated disease.

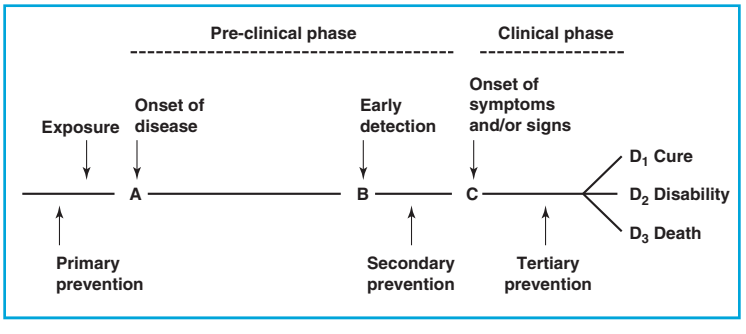

Figure 16.1.
Natural history of the disease 'cancer' and levels of prevention.

Implicit in this scheme is the notion that a disease evolves over time and that, as this occurs, pathological changes may become irreversible. The aim of prevention is to stop this progression.

There are various levels of prevention:

* *Primary prevention* is prevention of disease by reducing exposure of individuals to risk factors or by increasing their resistance to them, and thus avoiding the occurrence of event A.

* *Secondary prevention* (applied during the pre-clinical phase) is the early detection and treatment of disease. Screening activities are an important component of secondary prevention. In Figure 16.1, point B indicates the point in time at which the disease is first detectable by an appropriate screening test. For example, it might refer to the time at which a cancer mass reaches the min-

imum size that can be seen by X-ray examination. Thus, the distance from point B to C represents the 'detectable pre-clinical phase'. The location of point B varies markedly from one individual to another, and also depends on the screening technique used.

* *Tertiary prevention* (appropriate in the clinical phase) is the use of treatment and rehabilitation programmes to improve the outcome of illness among affected individuals.

In the rest of this chapter, we consider each of these levels of prevention in detail.

16.2 Primary prevention

The purpose of primary prevention is to limit the incidence of cancer by controlling exposure to risk factors or increasing individuals' resistance to them (e.g., by vaccination or chemoprevention). Clearly, the first step is to identify the relevant exposures and to assess their impact on the risk of the disease in the population.

16.2.1 How important is a particular exposure?

Relative and absolute measures of exposure effect

Much of epidemiology is concerned with identifying the risk factors for a disease, health problem or state of health. In assessing the strength of the association between a particular exposure and a particular outcome, we calculate measures known as *relative measures of effect*. As shown in Section 5.2.1, there are three types of relative measure (risk ratio, rate ratio and odds ratio) which are often collectively called relative risk.

Relative measures of effect provide answers to the question: *How many times more likely are people exposed to a putative risk factor to develop the outcome of interest relative to those unexposed, assuming that the only difference between the two groups is the exposure under study?* The magnitude of the relative risk is an important consideration (but not the only one—see Chapter 13) in establishing whether a particular exposure is a cause of the outcome of interest.

Once we have established that an exposure is *causally* associated with the outcome of interest, it is important to express the absolute magnitude of its impact on the occurrence of the disease in the exposed group (see Section 5.2.2). If we have information on the usual risk (or rate) of a particular disease in the absence of the exposure, as well as in its presence, we can determine the *excess risk* (also known as *attributable risk*) associated with the exposure.

Excess risk = risk (or rate) in the exposed – risk (or rate) in the unexposed

It is useful to express the excess risk in relation to the total risk (or rate) of the disease among those exposed to the factor under study. This measure is called *excess fraction* (also known as *excess risk percentage* or *attributable risk percentage*). It describes the proportion of disease in the *exposed group* which is attributable to the exposure.

Excess fraction (%) = 100 × (excess risk / risk (or rate) in the exposed)

Alternatively, it can be calculated by using the following formula:

Excess fraction (%) = 100 × (relative risk – 1) / relative risk

Example 16.1. *Suppose that a cohort study was conducted in the town of Minas Gerais (Brazil) to assess the relationship between cigarette smoking and lung cancer in men. We assume, for simplicity, that smokers and non-smokers were similar with respect to other risk factors for lung cancer such as age and occupational exposures. The results are shown in Table 16.1.*

	Smokers	Non-smokers	All
Number of cases	120	10	130
Person-years	54 545	50 000	104 545
Rate per 100 000 pyrs	220	20	124

Table 16.1.
Lung cancer incidence in smokers and non-smokers: hypothetical data.

In this study, the rate ratio is

220 per 100 000 pyrs / 20 per 100 000 pyrs = 11

and the excess risk associated with smoking (assuming causality) is

Excess risk = 220 per 100 000 pyrs – 20 per 100 000 pyrs = 200 per 100 000 pyrs

To assess what proportion 200 per 100 000 pyrs is of the rate among smokers *(220 per 100 000 pyrs), we can calculate the excess fraction:*

Excess fraction (%) = 100 × (200 per 100 000 pyrs / 220 per 100 000 pyrs) = 91%

This is the proportion of lung cancer cases in smokers *attributable to smoking.*

Excess fraction provides an answer to the question: *What is the proportion of new cases of disease in the exposed that can be attributed to exposure?* Another way of using this concept is to think of it as the decrease in the

incidence of a disease that would have been seen if the exposed had never been exposed. Thus, in Example 16.1, a maximum of 91% of lung cancer cases in smokers could theoretically have been prevented if they had never smoked.

If the exposure is *protective,* analogous measures can be calculated. They are usually called *risk reduction* (also known as *prevented risk*) and *prevented fraction* (also known as *prevented risk percentage*).

Risk reduction = risk (or rate) in the unexposed – risk (or rate) in the exposed

Prevented fraction (%) = 100 × (risk reduction / risk (or rate) in the unexposed)

Example 16.2. *A large randomized trial was carried out to assess the value of a smoking cessation programme (the intervention) in reducing the occurrence of lung cancer among smokers. By the end of the trial, the incidence of lung cancer was 155 per 100 000 pyrs among those who received the intervention and 240 per 100 000 pyrs among the controls. Thus, the maximum benefit achieved by the intervention was*

Risk reduction = 240 per 100 000 pyrs – 155 per 100 000 pyrs = 85 per 100 000 pyrs

Thus, 85 new cases of lung cancer per 100 000 pyrs were prevented by the smoking cessation programme.

Prevented fraction (%) = 100 × (85 per 100 000 pyrs/ 240 per 100 000 pyrs) = 35%

Thus 35% of the expected lung cancer cases among smokers were prevented by the smoking cessation programme.

Prevented fraction tends to be appreciably smaller than excess fraction (Example 16.2). This is because it is generally impossible to eliminate the exposure completely and, even if possible, the incidence of the disease in those who stop being exposed may never fall to the level in those who have never been exposed.

Calculation of excess risk (or risk reduction) requires information on the incidence of disease in the exposed and unexposed groups. This information is directly available in cohort and intervention studies. For case–control studies, however, it is not possible to calculate the excess risk using the formula given above, because incidence of disease among the exposed and unexposed groups is not known. We can still calculate excess fraction using the formula based on relative risk, which in these studies is estimated by the odds ratio. Alternative formulae can, however, be used in population-based case–control studies to calculate excess risk. These are presented in Appendix 16.1.

Measures of population impact

The measures of effect discussed so far compared the incidence of the disease in the *exposed* group with the incidence in the unexposed group. To assess the extra disease incidence in the *study population* as a whole that can be attributed to the exposure, we can calculate a measure called the *population excess risk* (also known as *population attributable risk*). This is defined as

Population excess risk = risk (or rate) in the population – risk (or rate) in the unexposed

or, similarly, as

Population excess risk = excess risk × proportion of the population exposed to the risk factor

Example 16.3. *Returning to the hypothetical study described in Example 16.1, the proportion of smokers in the whole cohort was 52%. If this 52% of the study population that smoked had never smoked, their incidence of lung cancer would have been reduced from 220 to 20 cases per 100 000 pyrs.*

Population excess risk = (220 per 100 000 pyrs – 20 per 100 000 pyrs) × 0.52 = 104 per 100 000 pyrs

Similarly, the population excess risk can be calculated by subtracting the rate in the unexposed group from the rate in the total study population. The rate in the total study population was 124 per 100 000 pyrs (Table 16.1). Thus,

Population excess risk = 124 per 100 000 pyrs – 20 per 100 000 pyrs = 104 per 100 000 pyrs

Thus, 104 cases of lung cancer per 100 000 pyrs could have been prevented in the whole study population *if none of the smokers had ever smoked.*

Analogously to the excess risk among exposed individuals, the *population excess risk* is a measure of the risk of disease in the *study population* which is attributable to an exposure (Example 16.3). We can express the population excess risk in relation to the total risk of the disease in the whole population. This measure is the *population excess fraction* (also known as *population attributable fraction*).

$$\text{Population excess fraction (\%)} = 100 \times \frac{\text{population excess risk}}{\text{rate (or risk) in the total population}}$$

Alternatively, it can be calculated by using the following formula:

$$\text{Population excess fraction (\%)} = 100 \times \frac{p_e\,(\text{relative risk} - 1)}{p_e\,(\text{relative risk} - 1) + 1}$$

where p_e represents the prevalence of exposure in the population under study.

Example 16.4. *In Example 16.3, the population excess fraction would be*

Population excess fraction (%) = 104 per 100 000 pyrs/124 per 100 000 pyrs = 84%.

This means that (assuming causality) approximately 84% of the lung cancer incidence in the study population is attributable to smoking. Thus, 84% of the lung cancer cases in this population would have been prevented if the smokers had never smoked.

Population excess fraction is an important measure. It provides an answer to the question: *What proportion (fraction) of the new cases of disease observed in the study population is attributable to exposure to a risk factor?* It therefore indicates what proportion of the disease experience in the population could be prevented if exposure to the risk factor had never occurred (Example 16.4).

Note that the excess fraction among the exposed is always greater than the population excess fraction, since the study population includes already some unexposed people who, obviously, cannot benefit from elimination of the exposure.

Sometimes it is useful to calculate the population excess fraction for a much larger population than the study population. For instance, public health planners are particularly interested in using data from epidemiological studies conducted in subgroups of the population to estimate the proportion of cases in a region or in a country that are attributable to a particular exposure (Example 16.5) . In this case, it is necessary to obtain data on the prevalence of exposure in these populations from other sources.

Table 16.2 shows how the population excess fraction varies in relation to the level of prevalence of the exposure in the population under study (p_e) and the magnitude of the relative risk. It is clear that the proportion of cases in a particular population that can be attributed to a particular exposure depends both on the magnitude of the relative risk and on the prevalence of the exposure in the population. For instance, tobacco smoking, with a relative risk of about 5, and occupational exposure to aromatic amines, with a relative risk of about 500, are implicated as causes of bladder cancer. Despite the fact that the relative risk is much smaller for smoking than for aromatic amines, the population excess fraction is sub-

Example 16.5. *Returning to the previous example, a recent household survey conducted in the region of Minas Gerais revealed that the prevalence of smoking among men was 35%. Thus, the proportion of lung cancer cases occurring among men in the whole region that can be attributed to smoking can be calculated as*

$$\text{Population excess fraction (\%)} = 100 \times \frac{0.35 \times (11 - 1)}{0.35 \times (11 - 1) + 1} = 78\%$$

Thus, in this hypothetical example, 78% of the lung cancers in the whole male population of Minas Gerais could be attributed to smoking. Note that this is lower than the value for the study population itself, the explanation being that the prevalence of smoking was lower in the whole male population of Minas Gerais (35%) than in the study population (52%).

stantially higher for smoking because this exposure is far commoner than exposure to aromatic amines. It has been estimated that the population excess fraction for smoking in England is 46% in men (Morrison *et al.*, 1984), whereas the population excess fraction for all occupational exposures (including exposure to aromatic amines) is only between 4 and 19% (Vineis & Simonato, 1986). Thus, a much larger number of bladder cancer cases would be prevented by eliminating smoking than by eliminating occupational exposures.

Prevalence of exposure (p_e) (%)	Relative risk		
	2	5	10
10	0.09	0.29	0.47
25	0.20	0.50	0.69
50	0.33	0.67	0.82
75	0.43	0.75	0.87
95	0.49	0.79	0.90

Table 16.2. Population excess fractions for different levels of prevalence of the exposure and various magnitudes of the relative risk.

These measures of population impact suffer from a number of limitations. Firstly, it has to be assumed that the risk factor is causally associated with the disease of interest. The criteria that may be used to assess whether an observed association is likely to be causal were discussed in Chapter 13. Secondly, it has to be assumed that there is no confounding or bias in the measures of incidence among exposed and unexposed groups. So far in our discussion we have, for simplicity, assumed that the exposed and unexposed groups were similar except for the exposure under study. This is rarely the case except in large randomized intervention trials. In our previous examples, for instance, we should have taken into account differences in the age distribution between smokers and nonsmokers. This can be done by using techniques similar to those described in Chapter 14 to calculate *adjusted measures*. We can then use these adjusted measures to calculate absolute measures of effect and measures of pop-

ulation impact using the same formulae as before. Thirdly, we must remember that estimates of relative risk are generally derived from case–control, cohort or intervention studies. These studies are often conducted in special subgroups of the population such as migrants, manual workers, etc. However, levels of exposure and intrinsic susceptibility in these subgroups may be quite different from those in the general population. It is therefore important that the extrapolation of data from these studies to other populations is undertaken with caution. For instance, many cohort studies are based purposely on groups with exposure to much higher levels than the general population (e.g., occupational cohorts) and the relative risks obtained from them should not be used as such to provide estimates of population excess fractions for other populations with much lower levels of exposure. This may be overcome if levels of exposure are properly measured (rather than just 'exposed' versus 'unexposed') and estimates of population excess fractions take them into account.

We can calculate the proportion of a particular cancer in a certain population that is caused by diet, by alcohol, by smoking, etc. These percentages may add up to more than 100%. This is because each individual calculation of population excess fraction does not take into account the fact that these risk factors interact with each other. For instance, in calculating the proportion of laryngeal cancer due to smoking, we ignore the fact that some of the cancers that occurred among smokers only occurred because they were also exposed to alcohol.

16.2.2 Role and evaluation of primary preventive measures

Once the risk factors have been identified and their impact in the population estimated, it is important to consider methods to either eliminate or reduce the exposure to them. Primary prevention involves two strategies that are often complementary. It can focus on the whole population with the aim of reducing average risk (the *population strategy*) or on people at high risk (the *high-risk individual strategy*). Although the high-risk individual strategy, which aims to protect susceptible individuals, is most efficient for the people at greatest risk of a specific cancer, these people may contribute little to the overall burden of the disease in the population. For example, organ transplant patients are particularly susceptible to non-melanoma skin cancer (Bouwes Bavinck *et al.*, 1991). The tumour tends to develop in highly sun-exposed areas of the body. Primary prevention campaigns for organ-transplanted patients involving reduction of sun exposure and sunscreen use are likely to be of great benefit to these patients, but will have little impact on the overall burden of disease in the population, because organ transplant patients represent a very small proportion of the population. In this situation, the population strategy or a combination of both strategies should be applied.

The major advantage of the population strategy is that it is likely to produce greater benefits at the population level and does not require identifi-

cation of high-risk individuals. Its main disadvantage is that it requires the participation of large groups of people to the benefit of relatively few. For example, adoption by a population of measures to reduce sun exposure may reduce the risk of skin cancer at a population level, but will be of little apparent benefit to most individuals, since the disease is rare even among those exposed. This phenomenon is called the *prevention paradox* (Rose, 1985).

Various approaches have been used to reduce or eliminate exposure to a particular risk factor, some examples of which are given in Box 16.1.

Box 16.1. Examples of approaches used to reduce or eliminate exposure to a hazard risk factor

- Health education on an individual or community basis (e.g., media campaigns promoting use of sunscreens).
- Regulation of carcinogens in occupational settings and in the environment (e.g., improvement of radiation protection).
- Price regulation (e.g., imposing taxes on cigarette and alcohol purchases).
- Advertising restrictions (e.g., banning of tobacco advertising or forcing the printing of health warnings on cigarette packages).
- Time and place restrictions on consumption (e.g., banning smoking in public places).

If specific preventive measures to reduce the incidence of a particular cancer have been adopted, it is essential to establish whether the effort has had any positive effect. Evaluation of primary preventive efforts at the population level is performed mainly in terms of monitoring changes in cancer incidence in relation to changes in exposure to risk factors. Thus, time trends in cancer incidence may be compared with temporal changes in exposure to a particular risk factor to show whether the desired effect is being achieved. This is illustrated in Figure 16.2, which shows trends in *per caput* consumption of cigarettes in the USA in relation to the timing of implementation of tobacco-control initiatives and important historical events, and trends in lung cancer mortality.

The following issues should be taken into account when interpreting incidence trends in relation to changes in exposure to risk factors. First, if the downward trend started long before the introduction of the preventive measure, it is difficult to attribute a recent decrease in incidence to the preventive measure under investigation. Second, given the long induction period of cancer, it may take many years or even decades before any effect of a preventive measure becomes apparent in incidence or mortality trends

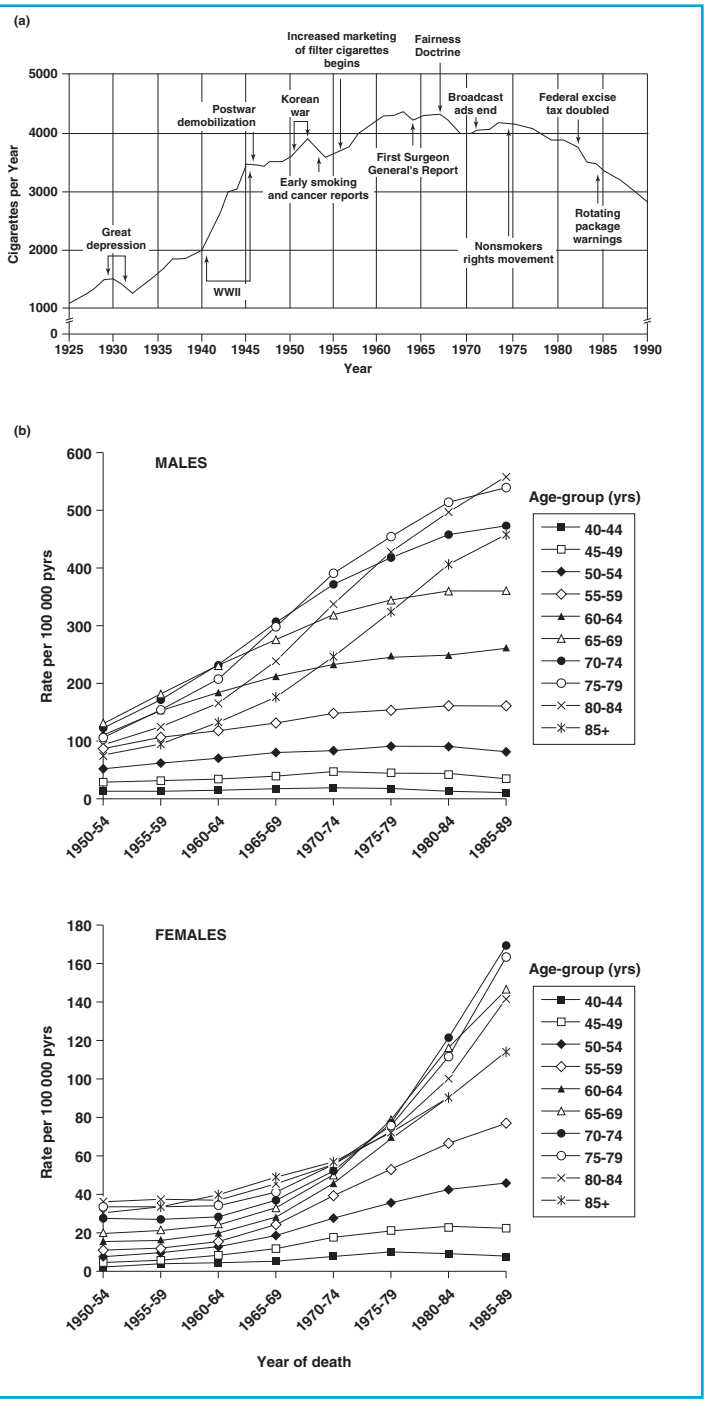

Figure 16.2.
(*a*) *Per caput* consumption of cigarettes among persons aged 18 years and over, United States of America, 1925–90 (reproduced from US Department of Health and Human Services, 1991) and (b) age-specific mortality trends from lung cancer, United States of America, 1950–89 (reproduced from Gilliland & Samet, 1994).

(except if the risk factor acts at late stages of carcinogenesis), as illustrated in Figure 16.2. Third, if the measures undertaken are directed to only a small fraction of the population (e.g., a region), the evaluation should be limited to the same sub-population, otherwise the effect may be missed. Fourth, when implementation has been confined to one area, comparisons of the changes in the intervention area versus 'control' areas may be possible.

Individual-based studies of subjects who have adopted potentially healthier habits or lifestyles are relatively few and were in the past confined mainly to the investigation of the risk of cancer in ex-smokers. Such studies show a marked decline in risk, which is related to the time since cessation of smoking, and they constitute the most powerful evidence for the effectiveness of stopping smoking in preventing cancer (Rose & Colwell, 1992; Gupta *et al.*, 1990). More recently, the potential of other changes in lifestyle (e.g., changes in diet) in cancer primary prevention have also been (or are currently being) assessed in large intervention trials (e.g. Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994; Hennekens *et al.*, 1996; Omenn *et al.*, 1996; Chlebowski & Grosvenor, 1994).

Individual-based studies have also helped to assess the effectiveness of preventive measures in the workplace. For instance, women first employed before 1925 in the watch dial-painting industry in the USA had greatly increased risks of mortality from bone cancer and from leukaemia and other haematological diseases, but risks declined for those employed in subsequent years (Table 16.3). The reduction in risk coincided with changes in work regulations in the industry, which included the prohibition of tipping or pointing of brushes between the lips in 1925–26 (Figure 16.3). These measures greatly reduced the exposure of the workers to radium.

Year of first employment	Bone cancer		Leukaemia and blood diseases	
	SMR (O/E)[b]	Observed no. of cases (*O*)	SMR (O/E)[b]	Observed no. of cases (*O*)
1915–19	233**	7	7.4*	2
1920–24	154**	20	3.3*	4
1925–29	10	1	1.0	1

[a] Data from Polednak *et al.* (1978)

[b] Expected numbers derived from cause-specific mortality rates for US white females

* $P < 0.05$; ** $P < 0.001$.

Table 16.3.
Mortality of women employed in US watch dial-painting industry and followed to end of 1975, by year of first employment.[a]

Other approaches to primary prevention are being evaluated, such as the use of mass vaccination campaigns (e.g., hepatitis B vaccine against liver cancer (Gambia Hepatitis Study Group, 1987)) and of chemoprevention (e.g., tamoxifen in prevention of breast cancer in high-risk women (Powles *et al.*, 1994)).

16.3 Secondary prevention

Secondary prevention refers to detection of cancer at an early stage, when treatment is more effective than at the time of usual diagnosis and treatment. With such measures it is possible to prevent the progression of the disease and its complications (including death).

16.3.1 Screening

Screening represents an important component of secondary prevention. It involves application of a relatively simple and inexpensive test to asymptomatic subjects in order to classify them as being likely or unlikely to have the disease which is the object of the screen. The positive cases can then be subjected to conventional diagnostic procedures and, if necessary, given appropriate treatment. Screening activities are based on the assumption that early detection and treatment will retard or stop the progression of established cases of disease, while later treatment is likely to be less effective. The ultimate objective of screening for a particular cancer is to reduce mortality from that disease among the subjects screened.

The concept of screening is not as straightforward as it may at first appear, however. Early treatment does not always improve prognosis and, even if it does, the true benefits of any type of screening have to be assessed in relation to its risks and costs and in relation to the benefits that may be derived from other public health activities. The final value of any screening programme can be established only by rigorous evaluation.

Any cancer screening activity requires (1) a suitable disease; (2) a suitable test and (3) a suitable screening programme.

Suitable disease

To be suitable for control by a programme of early detection and treatment, a disease must pass through a pre-clinical phase during which it is detectable (see Figure 16.1), and early treatment must offer some advantage over later treatment. Obviously, there is no point in screening for a

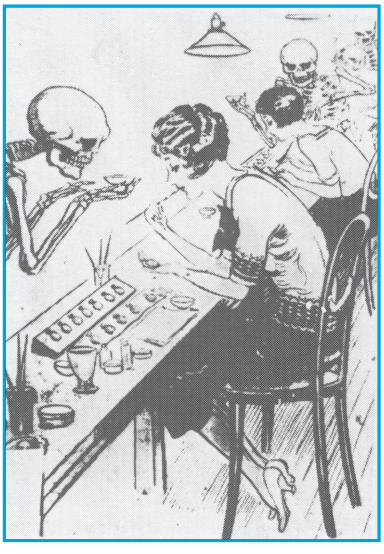

Figure 16.3.
New York newspaper cartoon alluding to the radium poisoning of watch dial painters

disease that cannot be detected before symptoms bring it to medical attention and, if early treatment is not especially helpful, there is no point in early detection.

Detectable pre-clinical phase

The pre-clinical phase of a cancer starts with the biological onset of the disease (point A in Figure 16.1). The disease then progresses and reaches a point at which it can be detected by the screening test (point B in Figure 16.1). From this point onwards, the pre-clinical phase of the disease is said to be 'detectable'. The starting point of this detectable pre-clinical phase depends partly on the characteristics of the individual and partly on the characteristics of the test being used. A test which can detect a very 'early' stage of the cancer is associated with a longer detectable pre-clinical phase than a test which can detect only more advanced lesions.

The proportion of a population that has detectable pre-clinical disease (its prevalence) is an important determinant of the utility of screening in controlling the disease. If the prevalence is very low, too few cases will be detected to justify the costs of the screening programme. At the time of initial screening, the prevalence of the pre-clinical phase is determined by its incidence and its average duration (recall the discussion on prevalence in Section 4.2). In subsequent screening examinations, however, the prevalence of the pre-clinical phase is determined mainly by its incidence, the duration being relatively unimportant if the interval between examinations is short. Therefore, the number of cases detected by the programme is greatest at the first screening examination, while the shorter the interval between examinations, the lower the number of cases detected per examination (and the higher the cost per case detected).

Early treatment

For screening to be of benefit, treatment given during the detectable pre-clinical phase must result in a lower mortality than therapy given after symptoms develop. For example, cancer of the uterine cervix develops slowly, taking perhaps more than a decade for the cancer cells, which are initially confined to the outer layer of the cervix, to progress to a phase of invasiveness. During this pre-invasive stage, the cancer is usually asymptomatic but can be detected by screening using the Papanicolaou (or Pap) smear test. The prognosis of the disease is much better if treatment begins during this stage than if the cancer has become invasive.

On the other hand, if early treatment makes no difference because the prognosis is equally good (or equally bad) whether treatment is begun before or after symptoms develop, the application of a screening test will be neither necessary nor effective and it may even be harmful (see below).

Relative burden of disease

Prevalence, incidence or mortality rates can be used to assess whether a cancer has sufficient public health importance to warrant instituting a

screening programme. Even if a disease is very rare but it is very serious and easily preventable, it may be worth screening for it. The final judgement will depend on the benefits, costs and cost/benefit ratio in relation to other competing health care needs.

Suitable test

For a screening programme to be successful, it must be directed at a suitable disease with a suitable test. In order to assess the suitability of a screening test, it is necessary to consider its validity and acceptability.

Validity

The preliminary assessment of a screening test should involve studies of its reliability, which is evaluated as intra- and inter-observer variation (see Section 2.6). Although even perfect reliability does not ensure validity, an

Box 16.2. Sensitivity, specificity and predictive value of a screening test

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

True positives = *a* False positives = *b*
True negatives = *d* False negatives = *c*

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

  Sensitivity = $a/(a + c)$

- The *specificity* of the screening test is the proportion of those classified as negatives by the gold standard who are correctly identified by the screening test:

  Specificity = $d/(b + d)$

- The *predictive value of a positive screening test result* represents the probability that someone with a positive screening test result really has the disease:

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

- The *predictive value of a negative screening test result* represents the probability that someone with a negative screening test result does not really have the disease:

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

unreliable test will not be sufficiently valid to be of use. On the other hand, a test that is highly valid must be highly reliable.

The validity of the screening test can be expressed by its sensitivity and specificity. These measure the ability of the test to identify correctly diseased and non-diseased people (see Box 16.2 and Section 2.6 for a review of the concepts). In the ideal situation, tests would be 100% sensitive and 100% specific. In this perfect situation, the distribution of the screening test results among individuals with the disease is completely separated from the distribution in healthy individuals. In Figure 16.4(*a*), everyone affected by the disease will have a test result which is above a chosen cut-off value located between these two distributions. Unfortunately, reality is more complex. In general, the distributions of the screening test results in the disease-free and diseased subjects overlap (Figure 16.4(*b*)). In these situations, the location of the cut-off value to classify screening test results as positive or negative is arbitrary. In Figure 16.4(*b*), some disease-affected subjects, having values below the chosen cut-off value, will be missed by the screening test, whereas some disease-free people will be wrongly classified as positive. Thus, there is generally a trade-off between the sensitivity and specificity of a given screening test; its ability to detect as many true cases as possible (high sensitivity) can only be increased at the expense of an increase in the number of individuals without the disease who will erroneously be classified as positive by the screening test (low specificity) and vice versa (see Section 2.6.1 for a numerical example of this).

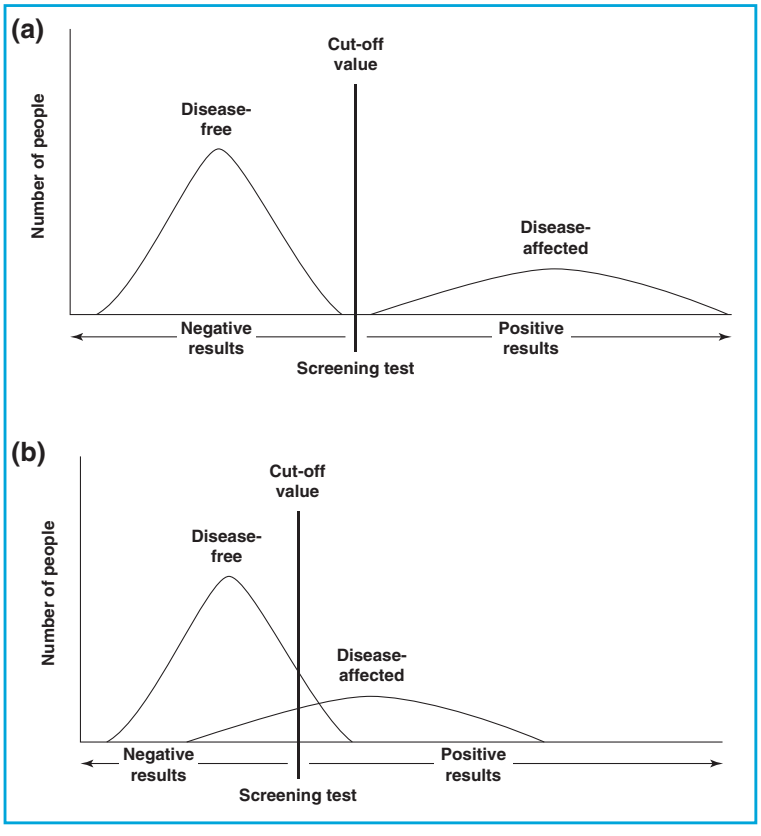

Figure 16.4.
Distribution of results from a screening test in disease-free and disease-affected individuals: (*a*) ideal distribution without overlap; (*b*) overlap of the distributions with the inevitable trade-off of sensitivity and specificity. In this example, where the screening test tends to give higher values for people with the disease than for those without it, moving the cut-off value to the left (i.e., lowering its value) increases sensitivity but decreases specificity; moving it to the right (i.e., increasing its value) increases specificity but decreases sensitivity.

Sensitivity is an indicator of the yield of cases (i.e., the number of truly diseased cases identified by the programme), whereas specificity is an indicator of the number of false-positive test results. Although one would like to detect all the subjects with the disease in a screening programme by using a test with a maximum sensitivity, such a policy might lead to an unacceptably low specificity, entailing high costs because of the need for further investigation of large numbers of false positives and a risk of poor motivation of subjects to participate in subsequent screening examinations. Hence, the choice of the cut-off point depends on the relative costs of false positives and false negatives.

In practice, however, it is difficult to estimate the sensitivity of the test, since it is not possible to apply a 'gold standard test' to the screened

population to find out the total number of diseased subjects ($a + c$ in Box 16.2). The screening test gives us only the value of a, that is, the number of persons who had a positive screening test and were confirmed to have the condition after further diagnostic evaluation. The usual approach to estimating sensitivity is to follow up subjects (usually for one year) having negative screening results, in order to observe how many cancers eventually develop among them. These '*interval*' cases are regarded as false negatives (c). The sensitivity of the screening test can then be calculated as usual. However, the value of this approach is limited since it is difficult to achieve complete follow-up and because some of the 'interval'cancers may have been true negatives at the time of the screening examination (i.e., very fast-growing tumours).

It is easier to estimate specificity if the screening is aimed at a rare condition such as cancer. Practically all those screened (N) are disease-free and thus N can be used to estimate the total number of people not affected by the condition ($b + d$ in Box 16.2). Since all screen-positive subjects are further investigated, the number of false positives (b in Box 16.2) is also known and, therefore, the number of true negatives (d in Box 16.2) can be calculated as $N - b$. Specificity can then be estimated as $(N - b)/N$.

Acceptability and costs

In addition to having adequate validity, a screening test should be low in cost, convenient, simple and as painless as possible, and should not cause any complications. Many screening tests meet these criteria—the Pap smear test for cervical precancerous lesions is a good example. In contrast, although sigmoidoscopic screening might lead to a reduction in mortality from colon cancer, it is questionable whether such a test would be acceptable because of the expense, the discomfort and the risk of bowel perforation.

Suitable screening programme

The organized application of early diagnosis and treatment activities in large groups is often designated as *mass screening* or *population screening*, and the set of procedures involved described as a *screening programme*.

A screening programme must encompass a diagnostic and a therapeutic component, because early detection that is not followed by treatment is useless for disease control. The diagnostic component includes the screening policy and the procedures for diagnostic evaluation of people with positive screening test results. The screening policy should specify precisely who is to be screened, at what age, at what frequency, with what test, etc., and it should be dynamic rather than fixed. The therapeutic component is the process by which confirmed cases are treated. It should also be dynamic and be regulated by strict universal procedures which offer the best current treatment to all identified cases.

A screening programme is a complex undertaking involving the application of a particular test to a particular population in a particular setting. Circumstances vary in different countries, and it should not be assumed that a format suitable for one country will apply to another without rigorous prior testing and evaluation. Box 16.3 lists the essential features of an organized screening programme.

Box 16.3. Essential features of an organized screening programme

- There is a clear definition of the target population.
- The individuals to be screened are identifiable (e.g., list with names and addresses of all eligible individuals in the target population).
- Measures are available to ensure high coverage and attendance (e.g., personal letter of invitation).
- There are adequate field facilities for collecting the screening material and adequate laboratory facilities to examine it.
- There is an organized quality-control programme to assess the screen material and its interpretation.
- Adequate facilities exist for diagnosis and appropriate treatment of confirmed neoplastic lesions and for the follow-up of treated individuals.
- There is a carefully designed referral system for management of any abnormality found, and for providing information about normal screening tests.
- Evaluation and monitoring of the total programme is organized so that it is possible to calculate incidence and mortality rates separately for those participating and those not participating in the programme, at the level of the total target population. Quality control of these epidemiological data should be established.

(modified from Hakama *et al.*, 1986)

16.3.2 Evaluation of screening programmes

Even after a disease is determined to be appropriate for screening and a valid test becomes available, it will remain unclear whether a widespread screening programme for that disease should be implemented in a particular population. It is therefore necessary to evaluate a potential screening programme to assess whether it is worth introducing it as a public health

measure to control a particular cancer. This involves consideration of two issues: first, whether the organization of the proposed programme is *feasible* and *cost-effective* (low cost per case detected), and second, whether it will be *effective in reducing the burden of the disease*. Both must be considered carefully. The implementation of a screening programme, no matter how cost-effective, will not be warranted if it does not accomplish its goal of reducing morbidity and mortality in the target population.

Process measures

The feasibility, acceptability and costs of a programme may be evaluated by *process measures*, which are related to the administrative and organizational aspects of the programme such as identification of the target population, number of persons examined, proportion of the target population examined, facilities for diagnosis and treatment in the health services, functioning of the referral system and its compliance, total costs, cost per case detected, etc. The major advantage of these process measures is that they are readily obtained and are helpful in monitoring the activity of the programme. Their main limitation is that they do not provide any indication of whether those screened have lower mortality from the cancer being targetted by the programme than those who were not screened.

A particularly useful process measure is the predictive value of a positive test. The predictive positive value (PPV) represents the proportion of persons found to have the disease in question after further diagnostic evaluation out of all those who were positive for the screening test ($a/(a+b)$ in Box 16.2). A high PPV suggests that a reasonably high proportion of the costs of a programme are in fact being spent for the detection of disease during its pre-clinical phase. A low PPV suggests that a high proportion of the costs are being wasted on the detection and diagnostic evaluation of false positives (people whose screening result is positive but did not appear to have the disease on subsequent diagnostic investigation). It is important to emphasize, however, that the PPV is a proportional measure; a high PPV might be obtained even if the frequency of case detection is unacceptably low. For instance, the PPV may be 80% indicating that 80% of those who screened positive were truly diseased. However, if only 10 subjects screen positive, the number of cases detected by the programme will be only 8! The main advantage of this measure is that it is available soon after the screening programme is initiated and, in contrast to sensitivity, no follow-up is necessary for it to be estimated.

The PPV of a screening test depends upon both the number of true positives a and the number of false positives b (see Box 16.2). Thus, it can be increased by either increasing the number of true positives or decreasing the number of false positives. The number of true positives may be increased by increasing the prevalence of detectable pre-clinical disease, for instance, by screening less frequently so as to maintain the prevalence of pre-clinical disease in the target population at a higher level. The number of false positives may be reduced by increasing the specificity of the

test, that is, by changing the criterion of positivity or by repeating the screening test after a positive test. A low PPV is more likely to be the result of poor specificity than of poor sensitivity. It is the specificity of a test that determines the number of false positives in people without the disease, who are the vast majority of people tested in virtually any programme. The sensitivity is less important for a rare disease because it operates on fewer people. By contrast, a small loss of specificity can lead to a large increase in the number of false positives, and a large loss of PPV.

Effectiveness in reducing mortality

The second, and definitive, aspect of evaluating a screening programme is whether it is effective in reducing morbidity and mortality from the disease being screened. Even if a screening programme will accurately and inexpensively identify large numbers of individuals with pre-clinical disease, it will have little public health value if early diagnosis and treatment do not have an impact on the ultimate outcome of those cases.

Obtaining an accurate estimate of a reduction in mortality requires a long-term follow-up of large populations. Consequently, *intermediate outcome measures* such as stage at diagnosis and survival (case-fatality) have been used which may be available in the early years of a screening programme. For example, in a successful screening programme, the stage distribution of the cancers detected should be shifted towards the less advanced stages and the risk of dying from cancer (case-fatality) should be lower for cases detected through screening than for symptom-diagnosed cases.

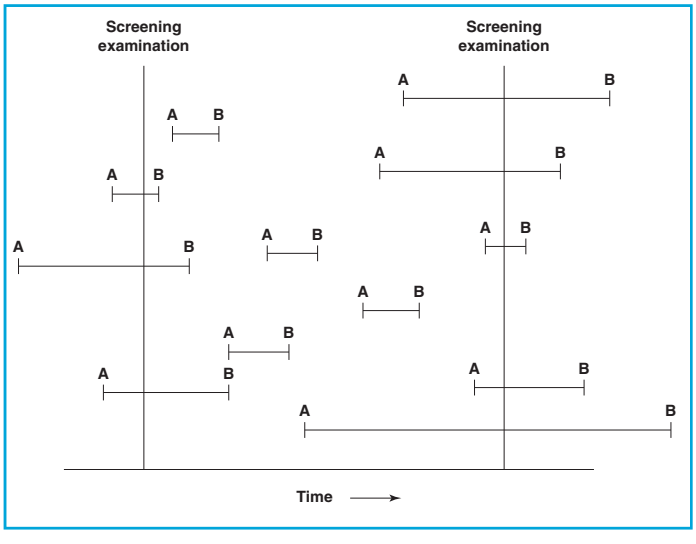

Figure 16.5.
Diagram illustrating length bias (see text). Each line represents a case from the point where it becomes detectable by the screening test (A) to the point where clinical symptoms or signs occur

There are, however, critical shortcomings associated with the use of these intermediate endpoints. Absence of a change in the parameters may mean that the screening is not successful, but they do not provide an adequate measure of evaluation because they suffer from a number of biases, namely, length bias, lead-time bias and overdiagnosis bias.

(*a*) *Length bias.* Length bias refers to the phenomenon occurring when cases detected by a screening programme are not a random sample from the general distribution of all cases of pre-clinical disease in the screened population. This is likely to happen when screening tests are applied at moderately long intervals (say once every 2–5 years), so that cases with a long pre-clinical phase are more likely to be detected than those with faster-growing tumours (Figure 16.5). Hence, the cases detected by screening may be those with lesions having a more favourable prognosis, while cases with similar onset date but more rapid disease progression are detected by clinical symptoms. The resulting length bias could lead to an erro-

neous conclusion that screening was beneficial when, in fact, observed differences in survival (case-fatality) were a result merely of the detection of less rapidly fatal cases through screening.

(*b*) *Lead-time bias*. If an individual participates in a screening programme and has disease detected earlier than it would have been in the absence of screening, the amount of time by which diagnosis is advanced as a result of screening is the *lead time*. Since screening is applied to asymptomatic individuals, by definition every case detected by screening will have had its diagnosis advanced by some amount of time. Whether the lead time is a matter of days, months, or years, however, will vary according to the disease, the individual, and the screening procedure. Cases progressing rapidly from pre-clinical to clinical disease will gain less lead time from screening than those that develop slowly, with a long pre-clinical phase. The amount of lead time will also depend on how soon the screening is performed after the pre-clinical phase becomes detectable. Because of the lead-time phenomenon, the point of diagnosis is advanced in time and survival as measured from diagnosis is automatically lengthened for cases detected by screening, even if total length of life is not increased. This is referred as lead-time bias.

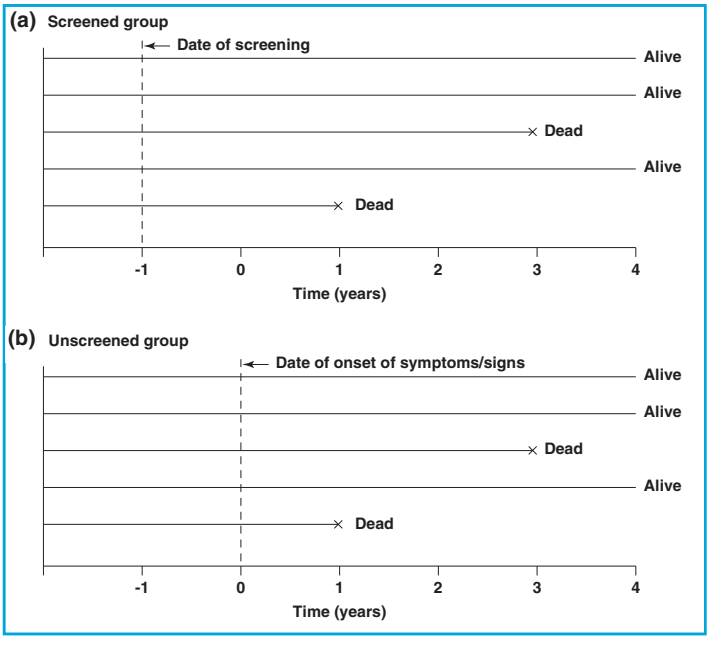

Figure 16.6.
Diagram illustrating lead-time bias. The lines represent the survival experience of each cancer case in the screened and unscreened groups (see text).

Suppose that 100 individuals were screened for a particular cancer for which there is no effective treatment. On average, the test succeeds in identifying the cancer one year before it becomes clinically evident. A similar unscreened group of 100 patients was also assembled. These two groups were followed up for five years and five persons were detected as having this cancer in each of them.

Let us examine the survival experience of the five cases in each of the groups. The course of their illness is shown in Figure 16.6. The 1.5-year survival for the screened group is 100% (all of them were still alive 1.5 year after being screened), whereas the 1.5-year survival for the unscreened group was only 80% (one case died one year after the onset of symptoms), even though the two groups have the same duration of survival (given our initial assumption of no effective treatment).

The problem with this analysis is that the starting point for monitoring survival is different between the screened and unscreened cases, always to the apparent detriment of the cases detected without screening. The appropriate approach is to compare the mortality experience of the 100 screened people with the mortality experience of the 100 unscreened peo-

ple *from the time of screening*. In the above example, the mortality rate in the screened group is two deaths in 496 person-years (98 persons × 5 years, plus 1 person × 4 years, plus 1 person × 2 years). The rate is the same in the unscreened group, since the number of person-years, counted from the time the screening would have taken place had it been done, is identical to that for the screened group.

There are two ways in which the effect of lead time on the evaluation of the efficacy of a screening programme can be taken into account. The first is to compare not the length of survival from diagnosis to death, but rather the mortality rates in the screened and unscreened groups (as done in the above example). Alternatively, if the lead time for a given disease can be estimated, it can be taken into account, allowing comparison of the survival experience of screen- and symptom-detected cases. For example, the average lead time for breast cancer has been estimated as approximately one year (Shapiro *et al.*, 1974). Thus, to evaluate the efficacy of a breast cancer screening programme, the two-, three-, four-, five- and six-year survival risks of the screened cases should be compared, respectively, with the one-, two-, three-, four- and five-year survival risks of the unscreened cases. However, determinations of lead time are difficult to carry out and cannot be generalized, since they depend on the ability of the screening procedure to detect pre-clinical conditions.

(*c*) *Overdiagnosis*. It is possible that many of the lesions detected by the screening programme would never have led to invasive cancer and death. These lesions are known as 'pseudo-cancer'. Thus, the true benefit of identifying pre-clinical lesions through screening may be much smaller than is perceived.

In short, although intermediate outcome measures, such as stage distribution and case fatality (survival), may appear to be suitable as surrogate endpoints in a screening programme, they are subject to lead time bias, length bias and overdiagnosis bias. Thus, the ultimate outcome measure which should be evaluated in screening programmes aimed at detecting early cancer (e.g., breast and colon cancer screening) is reduction in mortality. When screening is aimed at detecting both pre-cancerous conditions and early cancers (e.g., cervical cancer screening), reduction in the incidence of invasive cancer and reduction in mortality are suitable outcome measures. This implies that any screening programme should be planned in such a way that its evaluation in terms of change in mortality (and incidence) in the total target population is possible.

An illustration of how intermediate outcomes may be misleading is given in Example 16.6. In this example, intermediate outcomes seemed to indicate that the use of chest X-ray and cytology was effective in lung cancer screening. However, no reduction in mortality was observed. Similar results have consistently been found in all randomized trials that have addressed this issue.

Example 16.6. *A total of 6364 cigarette-smoking males aged 40–64 years were randomized into an intervention group which received six-monthly screening by chest X-ray and sputum cytology during three years, and a control group which received a single examination at the end of the third year. Lung cancer cases detected by screening were identified at an earlier stage, were more often resectable, and had a significantly better survival than symptom-detected cases. There was, however, no significant difference in mortality between the intervention and control groups (Kubik* et al.*, 1990).*

Studies to evaluate the effectiveness of a screening programme

Intervention studies

The randomized trial is the best study design for evaluating the effectiveness of a screening programme, because it provides the opportunity for a rigorous experimental evaluation. When the sample size is sufficiently large, control of confounding is virtually assured by the process of randomization. Patient self-selection or volunteer bias, which is problematic for the comparison of screened and unscreened groups in observational studies, cannot influence the validity of the results of randomized trials, since the screening programme is allocated at random by the investigators after individuals have agreed to participate in the trial.

There are various problems with randomized trials, however. First, there can be contamination of the control group (awareness of the screening programme may lead subjects in the control group to seek screening). Second, a large number of subjects may be required in screening trials for diseases with low incidence rate, such as most cancers, and/or if the trial is designed to show small benefits (as in Example 16.7). Third, it may be unacceptable to randomize some subjects to be non-screened if a screening programme has already been introduced despite the lack of experimental evidence (e.g., screening for cervical cancer).

Example 16.7. *A randomized controlled trial was set up in Sweden in 1977 to assess the efficacy of mass screening with single-view mammography in reducing mortality from breast cancer. A total of 162 981 women aged 40 years or more and living in the counties of Kopparberg and Östergötland were randomized to either be or not be offered screening every 2 or 3 years, depending on age. The results to the end of 1984 showed that among women aged 40–74 years at the time of entry, there was a 31% reduction in mortality from breast cancer in the group invited for screening (Tabár* et al.*, 1985).*

Observational studies

While randomized trials can provide the best and most valid evidence concerning the efficacy of a screening programme, as with the evaluation of etiological hypotheses, most evidence on the effects of screening pro-

grammes will come from the non-experimental study designs because of issues of costs, ethics and feasibility. Especially in the numerous situations where randomized trials are not possible, such as with well established procedures like the Pap smear, observational approaches can provide useful and necessary information. Interpretation of the results from these studies is less straightforward, however.

Ecological studies have been used to examine trends in disease rates in relation to screening frequencies within a population, or to compare the relationship between the frequencies of screening and disease rates for different populations (as in Example 16.8). Such studies can be useful in suggesting that a relationship exists between screening and a decline in morbidity or mortality, but the inherent limitations of ecological studies must be borne in mind. First, since the information from such studies concerns populations rather than individuals, it is not possible to establish that those experiencing the decreased mortality are in fact the same persons who were screened (the *ecological fallacy*; see Section 11.2.2). Moreover, such studies cannot allow for control of potential confounding factors. Finally, the measure of screening frequency employed is usually an average value for the population, so that it is not possible to determine an optimal screening strategy for an individual. Thus, ecological studies may suggest the

Example 16.8. *To assess whether declines in cervical cancer mortality in the USA were related to screening, the change in cervical cancer mortality between 1950–54 and 1965–69 in the various states was examined in relation to the proportion of women screened in each one. The results are shown in Figure 16.7. There was a positive correlation between the magnitude of the decrease in mortality rates and the screening effort in each state. The positive relationship becomes more evident if the less populous states are excluded from the analysis (correlation coefficient (r) = 0.60; P<0.0005 (Cramer, 1974). These results were consistent with a beneficial effect of cervical cytological screening. However this relationship may be confounded by other factors such as socioeconomic changes.*

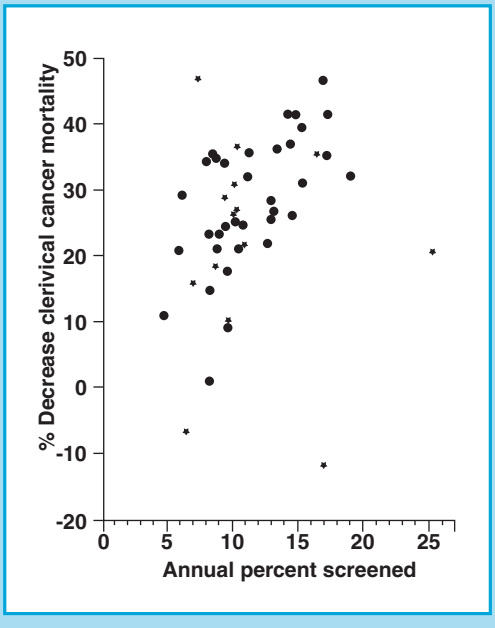

Figure 16.7.
Change in cervical cancer mortality between 1950–54 and 1965–69 in relation to estimated average annual percentage of women aged 19 years and over screened during 1953–68, by state of the USA. The stars indicate the less populous states (i.e., accumulated female population aged 19 years or over of less than five million during the study period) (reproduced, by permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc., from Cramer, 1974).

possibility of a benefit of a screening programme, but they cannot test that hypothesis.

Cohort studies require long-term follow-up of screened and unscreened subjects. However, in interpreting the results of such studies, the potential effects of self-selection of participants must be taken into account (Example 16.9).

Example 16.9. *To examine the protective effect of Pap-smear screening for cervical cancer, a cohort study was conducted in two counties in Sweden (Uppsala and Gävleborg) where organized cytological screening was introduced in 1967 and 1972, respectively. A total of 386 990 women resident at any time during 1968 to 1992 in these two counties were identified through the Population Register and enrolled into the study. Each woman's screening history was ascertained from computerized registers of Pap smears taken in the area and record-linkages allowed complete follow-up with regard to cancer incidence, out-migration and survival through to 1992. A total of 938 newly diagnosed cases of squamous cell cervical cancer occurred during the follow-up of this cohort. Women who were ever screened were found to have about half the risk of those never screened (rate ratio = 0.55; 95% confidence interval 0.51–0.61) (Sparén, 1996).*

In cohort studies such as that described in Example 16.9, the people undergoing screening are not chosen randomly and individuals who choose to be screened may differ both from those who refuse screening and from the population at large (selection bias). These volunteers may have very different prognoses compared with their unscreened counterparts. In general, volunteers tend to have better health and lower mortality rates than the general population and are more likely to adhere to prescribed medical regimens. On the other hand, those who volunteer for a screening programme may represent the 'worried well', that is, asymptomatic individuals who are at higher risk of developing the disease because of medical or family history, or any number of lifestyle characteristics. Such individuals may have an increased risk of mortality regardless of the efficacy of the screening programme. The direction of the potential selection bias may be difficult to predict and the magnitude of such effects even more difficult to quantify.

Case–control studies of screening involve comparison of the screening histories of subjects who do, or do not, exhibit the outcome which screening aims to prevent (death from cancer or incidence of invasive disease). Although case–control studies are increasingly used to evaluate screening programmes, they cannot replace experimental studies because they are liable to confounding and bias (for instance, cases may differ from controls in their ability to recall past screening). However, once a form of screening is widespread, case–control studies may make use of existing records so that recall bias should not arise. The first study of this kind

compared the history of Pap smear screening in 212 hospital cases of invasive cervical cancer with age-matched neighbourhood controls (Clarke & Anderson, 1979). Fewer cases than controls had received a Pap smear during the five years before the year of diagnosis. The risk of invasive cervical cancer among women who had not had a Pap smear was about three times that of women who were screened, after controlling for socioeconomic status. The authors attempted to examine the impact of potential recall bias by comparing the data obtained through a sample of personal interviews with data from physicians' records. There was no evidence that the information obtained during the interviews was affected by recall bias.

***Example 16.10**. In Cali (Colombia), screening for cervical cancer has been offered to all sexually active women and routinely performed in pre-natal clinics since the late 1960s. To evaluate the role of Pap smear in preventing invasive carcinoma of the cervix in this population, a case–control study was carried out. A total of 204 cases with newly diagnosed invasive cervical cancer during the years 1977–81 were identified through the Cali population-based cancer registry and successfully interviewed. For each case, a neighbourhood control matched to year of birth of the case ± 2 years was selected. Cases and controls were interviewed about history of Pap smears performed for screening purposes during the period 12–72 months before the date on which the case was diagnosed. Examinations performed within 12 months of diagnosis were ignored because they were likely to be symptom- or disease-related. For each control, the inquiry covered the same calendar time interval as that of the matching case, as determined by her date of diagnosis. The risk of developing invasive carcinoma was 10 times greater in unscreened than in screened women (Aristizabal* et al.*, 1984).*

The use of case–control studies to evaluate screening programmes raises some special methodological issues. Tests done for diagnostic rather than screening purposes should not be considered and, for the controls, the screening history should be restricted up to the time of diagnosis of the case (as in Example 16.10) to ensure that cases and controls are fully comparable with respect to period of exposure to screening.

Case–control studies can be set up as an integral part of established screening programmes, in order to assess the screening policy (e.g., the age at which screening should be initiated or stopped, or the optimal frequency of screening).

16.3.3 Targeting high-risk groups

One way of reducing the costs of a screening programme is to target the screening towards groups of individuals at higher than average risk of developing the disease of interest. Most cancer screening programmes are limited to certain age-groups. For instance, it is not worth screening women under age 40 for breast cancer, because very few cases occur at these ages.

Screening programmes can also be targeted exclusively to high-risk groups defined on the basis of factors such as family history, medical history or occupation. For instance, targeting breast cancer screening programmes exclusively to women with a positive family history would increase the proportion of cases detected among screened women, but the large majority of cases in the population would be missed since they occur among people without a family history of this disease. Thus, restricting a screening programme to selected high-risk groups is useful only if a substantial fraction of all the cases in a population occur in these high-risk groups.

16.4 Tertiary prevention

Tertiary prevention consists of alleviation of disability resulting from disease in order to improve the outcome of illness among affected individuals. It includes not only the treatment itself, but also all rehabilitation attempts to restore an affected individual to a useful, satisfying, and, where possible, self-sufficient role in society.

Randomized clinical trials are the only acceptable method to evaluate cancer treatments (see Chapter 7). However, data from population-based cancer registries may provide a more representative picture for evaluating comprehensive cancer care in a particular population, since they will include all cancer cases in the population regardless of the treatment they might or might not have received. These issues are further discussed in Section 17.6.2.

Box 16.4. Key issues

- Epidemiology is a key discipline of public health which provides the scientific background for formulation of policies aimed at preventing the development of disease in healthy persons.

- There are three levels of cancer prevention:

  (a) Primary prevention aimed at preventing the onset of the disease either by reducing exposure to risk factors or by increasing the individuals' resistance to them. Measures of population impact are very useful in helping to identify exposures that are potentially responsible for large numbers of cases of a particular cancer in the population.

  (b) Secondary prevention aimed at reducing mortality from a particular cancer through early detection and treatment. Screening programmes are an important part of secondary prevention.

  (c) Tertiary prevention aimed at improving the prognosis and quality of life of affected individuals by offering them the best available treatment and rehabilitation programme.

Further reading

* A comprehensive discussion of cancer screening programmes is given in Cole & Morrison (1978).

Box 16.4. (Cont.)

- *Screening* involves the use of a simple and inexpensive test to detect early stages of cancer at which treatment is more effective than at the time of usual diagnosis. Mass screening programmes should only be directed towards the control of cancers for which there is an effective treatment that will reduce mortality, if applied at early stages. There should also be valid, inexpensive and acceptable tests for the detection of the cancer at early stages.

- The performance of a screening test in terms of its acceptability, feasibility and costs can be monitored by *process measures* related to the administrative and organizational aspects of the programme (e.g., proportion of the target population examined, functioning of the referral system, cost per case detected, etc.). Predictive value of a positive test is a particularly useful measure because it provides an indication of whether most of the effort of the programme is being used to identify cases at an early stage or whether they are mainly wasted on the evaluation of false positives. Although process measures are useful for monitoring the activity of the programme they do *not* indicate whether those screened will have lower mortality than those not screened.

- The ultimate outcome measure to be used in evaluating the *effectiveness* of a screening programme aimed at detecting early cancer (e.g. mammography) is reduction in mortality. When the programme is aimed at detecting both pre-cancerous conditions and early cancer (e.g., Pap-smear screening), reduction in the incidence of invasive cancer and reduction in mortality are suitable outcome measures. *Intermediate outcome measures* such as stage distribution and case-fatality (survival) have also been used, but although they give an indication of whether the programme is likely to be effective, they are subject to length bias, lead time bias and overdiagnosis bias.

- The effectiveness of a screening programme should ideally be assessed by conducting a randomized intervention trial. In practice, most of the evidence on the effects of screening programmes comes from observational studies.

Appendix 16.1

Calculation of absolute measures of exposure effect and measures of population impact in case–control studies

In *population-based case–control studies* in which the incidence rate in the total population of interest is known and the distribution of exposure among the controls is assumed to be representative of the whole population, these parameters can be used to estimate incidence rates in the exposed and unexposed groups.

A population contains a mix of exposed and unexposed people. Thus, the overall incidence rate (r) of the disease in a population is equal to the weighted average of the incidence rates in its exposed (r_1) and unexposed (r_0) groups, the weights being the proportions of individuals in each group. Suppose that a proportion p_e of the population is exposed to the factor under study. Thus, the proportion of unexposed people in that population is equal to $(1 - p_e)$. Hence, the rate in the population will be

$$r = r_1 p_e + r_0 (1 - p_e)$$

Since the relative risk (estimated by the odds ratio (OR) in case–control studies) is the ratio of the incidence rates among the exposed and unexposed, the incidence rate among the exposed (r_1) members of a population is equal to the relative risk times the rate in the unexposed ($\text{OR} \times r_0$). Hence,

$$r = (r_0 \times \text{OR} \times p_e) + r_0 (1 - p_e)$$

$$= r_0 ((\text{OR} \times p_e) + (1 - p_e))$$

$$r_0 = \frac{r}{(\text{OR} \times p_e) + (1 - p_e)}$$

Once the incidence rate among the unexposed is determined, it can be multiplied by the odds ratio to provide an estimate of the incidence among the exposed. Given these two incidence rates (r_1 and r_0), the *excess risk* and the *excess fraction* can then be calculated as usual.

Example A16.1. *In a hypothetical population-based case–control study conducted in London, cases with lung cancer were nine times more likely to have smoked cigarettes regularly in the past five years than men without lung cancer. The population lung cancer incidence rate in London was 40 per 100 000 pyrs and the proportion of smokers among the controls 60%. Thus,*

$$r_0 = \frac{40 \text{ per } 100\,000 \text{ pyrs}}{(9 \times 0.6) + (1 - 0.6)} = 6.9 \text{ per } 100\,000 \text{ pyrs}$$

$$r_1 = 9 \times 6.9 \text{ per } 100\,000 \text{ pyrs} = 62.1 \text{ per } 100\,000 \text{ pyrs}$$

$$\text{Excess risk} = 62.1 \text{ per } 100\,000 \text{ pyrs} - 6.9 \text{ per } 100\,000 \text{ pyrs} = 55.2 \text{ per } 100\,000 \text{ pyrs}$$

$$\text{Excess fraction (\%)} = 100 \times (55.2 \text{ per } 100\,000 \text{ pyrs} / 62.1 \text{ per } 100\,000 \text{ pyrs}) = 89\%$$

Note that the excess fraction could also have been calculated by using the formula given in Section 16.2.1:

$$\text{Excess fraction (\%)} = 100 \times \frac{(OR - 1)}{OR}$$

$$\text{Excess fraction (\%)} = 100 \times \frac{(9 - 1)}{9} = 89\%$$

The *population excess risk* and the *population excess fraction* can then be determined as

$$\text{Population excess risk} = \text{excess risk} \times \text{proportion of the population exposed to the factor } (p_e)$$

$$\text{Population excess fraction (\%)} = 100 \times \frac{\text{population excess risk}}{\text{rate in the total population } (r)}$$

The *population excess fraction* can also be calculated by using the formula given in Section 16.2.1:

$$\text{Population excess fraction (\%)} = 100 \times \frac{p_e\,(OR - 1)}{p_e\,(OR - 1) + 1}$$

Example A16.2. *In the above example, the population excess risk and the population excess fraction can be calculated as follows:*

Population excess risk = 55.2 per 100 000 pyrs × 0.6 = 33.1 per 100 000 pyrs

$$\textit{Population excess fraction (\%)} = 100 \times \frac{0.6\,(9-1)}{0.6\,(9-1)+1} = 83\%$$

Thus, 83% of the lung cancer cases in the whole population of London could be attributed to smoking.

In *hospital-based case–control studies*, it is not possible to calculate the excess risk or the population excess risk, since incidence rates in the exposed and unexposed cannot be estimated. However, the following formulae can be used to calculate *excess fraction* and *population excess fraction:*

$$\text{Excess fraction (\%)} = 100 \times \frac{(\text{OR}-1)}{\text{OR}}$$

$$\text{Population excess fraction (\%)} = 100 \times \frac{p_e\,(\text{OR}-1)}{p_e\,(\text{OR}-1)+1}$$