Chapter 5

Effectiveness of screening

Has screening been implemented in accordance with the results of screening trials?

National screening programmes were introduced in a number of countries, including Sweden, The Netherlands and the United Kingdom, following randomized controlled trials, and the experience from the trials was used to decide how screening should be implemented. A number of trials conducted after the start of those programmes concentrated on more specific issues and also influenced national programmes.

There are now many national and regional population-based screening programmes, with various characteristics (see Tables 9–11, pp. 49–53). Ballard-Barbash *et al.* (1999) reviewed breast cancer screening programmes in 21 countries and divided them into three groups:

- those with national, government-supported, centrally organized health care systems, which have highly organized screening programmes that are distinct from the delivery of general medical care;
- those with government-supported programmes that are organized more locally; and
- those with government-supported programmes, where breast cancer screening is conducted within the context of general medical care.

These distinctions were used to focus the discussion below, although examples from the first group were used mainly.

The screening process can be a continuum, from organized national programmes through to opportunistic screening in which no 'programme' as such effectively exists (see Chapter 3). Organized programmes cover a defined population, and women in a specific age range are invited at regular intervals. These programmes are based on information on all women in a geographical area, which allows tracking of both those who attend and those who do not. Organized programmes can be evaluated by the techniques described below. Opportunistic screening is that conducted after a request by a physician or a woman rather than by specified routine call and recall. In practice, many programmes have elements of both organized and opportunistic screening. By its nature, opportunistic screening is less amenable to quality control and the detailed evaluation specified for organized screening programmes; the evaluation may therefore be limited to technical considerations.

Methods of invitation

For the majority of the randomized trials, population registers were obtained or compiled before randomization, from electoral or health authority registers (Tabár *et al.*, 1985; Roberts *et al.*, 1990) or from private health plan membership lists (Shapiro *et al.*, 1967), and women in the intervention group were sent personal invitations. An exception is the Canadian trials (Miller *et al.*, 1991b), in which women volunteered after media publicity. The advantages of a population-based approach with individual invitations is that the whole eligible population and range of socioeconomic groups are identified. In most countries where this is feasible, such an approach was adopted for national screening programmes. In many countries, existing population registers were used, whereas in others (e.g. Ireland) registers had to be compiled from various sources. In a survey of screening in 21 countries, personal invitation was used as the only means for recruitment in five, and this was the commonest method in a further nine (Ballard-Barbash *et al.*, 1999). In the USA, referral by a physician was the commonest method of recruitment. Media publicity was used in a number of countries, but it was the commonest method of recruitment in only two.

In countries or regions where recruitment is not organized by invitation, it may be difficult to encourage certain groups to attend for screening, and specific interventions may be needed (Whitman *et al.*, 1991; Paskett *et al.*, 1999). In addition, it may prove more difficult to ensure that such women attend for further investigation whenrecommended or return for routine re-screening (Segura *et al.*, 2000). In the province of Florence, Italy, a self-referral policy resulted in only 10.2% of the target population having a mammogram (Giorgi *et al.,* 1994).

Screening processes

The screening process is described comprehensively in Chapter 3. It should be emphasized that screening comprises a series of elements, starting from identification of the target population, through invitation of the woman and the point at which a women has a negative screening result or breast cancer is detected and treated, and the outcome is evaluated. Detailed information on the screening trials that have been performed is given in Chapters 1 and 4, including those with conventional mammography. This chapter addresses how the trials have influenced the implementation of screening. In most of the randomized trials, the further assessment of women with abnormalities detected on screening mammograms was also described. The extent to which this is done in population screening programmes depends partly on the health care system.

Age range

The most convincing evidence from randomized controlled trials of the benefit of screening was for women aged 50–69 at entry to the trial. Few results are available for women aged ≥ 70. With further results and increased follow-up, a benefit is becoming apparent from some trials for women younger than 50 at entry, although it remains unclear to what extent this is due to screening after the age of 50 (de Koning *et al.,* 1995b; Fletcher, 1997; National Institutes of Health, 1997; see also Chapter 4).

All countries in which population screening is conducted (either nationally or sub-nationally) include women aged 50–64, and a number include women up to the age of 69. Concern about lower attendance of older women led to an initial upper limit for invitation of 64 in the United Kingdom, but on the basis of the results of recent studies (Moss *et al.*, 2001), the upper limit is to be increased to 70, and that in The Netherlands has been increased to 74. In some countries, including the USA, there is no recommended upper limit. As stated above, little evidence is available from the trials about the benefit of screening women > 70.

If screening women younger than 50 is to be effective, the interval will probably need to be < 2 years. A randomized controlled trial of women aged 40–41 at entry given annual mammography is being conducted in the United Kingdom, and a report will become available within the next few years (Moss, 1999). Breast cancer is less common among younger women, and, while more life-years will potentially be gained, cost–effectiveness will be a key consideration for this age group. At present, in most countries with screening programmes, routine screening of women under 50 is not recommended (National Institutes of Health, 1997; Ballard-Barbash *et al.*, 1999).

Screening interval

Most of the randomized controlled trials involved a 1- or 2-year screening interval. In the Swedish Two-county study, a 24-month interval was used for women aged 40–49 and a 33-month interval for those aged 50–75 at entry. Tabár *et al.* (1989) recommended that the interval should be no more than 18 months for women aged 40–49 and no more than 2 years for women aged ≥ 50, on the basis of the results of randomized controlled trials.

Many countries have adopted a 2-yearly screening interval because of the high interval cancer rates seen in the third year in trials. The United Kingdom is unusual in opting for a 3-year interval for women aged 50–64, but the programme is restricted by budget. The Microsimulation Screening Analysis (MISCAN) model (see p. 128) predicts that substantial reductions in mortality would follow from extending the age range of women screened to 69 or reducing the screening interval to 2 years and suggests that the difference between these two policies would be so small that, depending on the outcome measure considered, either could be used (Boer *et al.*, 1998). The results of Markov-chain models of breast tumour progression to determine the optimal screening interval (Duffy *et al.*, 1997), with the data from the Swedish trials, suggest that the screening interval is critical for women aged 40–49 but less so for older women. For women aged 40–49, a 3-year interval would result in little reduction in mortality (4%), but annual screening would result in a 36% reduction. For women aged 60–69, the predicted reductions in mortality with 3-yearly, 2-yearly and annual screening are predicted to be 34%, 39% and 44%, respectively.

Mammography

Number of mammographic views

In mammography, a single mediolateral–oblique view with or without a second cranio-caudal view is usually used. Single-view mammography was used in the Swedish Two-county and Stockholm trials, while two-view mammography was used in the Malmö, Göteborg, Health Insurance Plan and Canadian trials and in the first screen in Edinburgh. The results of a meta-analysis of these trials (Kerlikowske *et al.*, 1995) suggested that there was no difference in the reduction in breast cancer mortality between screening with one view or two views; however, the trials were not designed to answer this question.

In most national screening programmes, two views are used for the prevalence screen; in some, two views are used for all screens. In The Netherlands, two views are used for prevalence screens and about 20% of all subsequent screens if indicated (Fracheboud *et al.*, 1998). In the United Kingdom, a single view was recommended for all screens; in 1995, however, on the basis of the results of the randomized controlled trial (Wald *et al.*, 1995), the country changed its policy to two views for prevalence screens. It recently changed the policy again, to recommend two views for all screens by 2003. In a national screening programme, film readers will, at least initially, be less skilled than the experts involved in randomized controlled trials. Experience in the United Kingdom showed a 32% increase in the detection of small invasive cancers when the programme changed from one view to two views at prevalence screens (Blanks *et al.*, 1998) and also showed that two views were beneficial at incidence screens (Given-Wilson & Blanks, 1999).

Double reading

Deciding how many film readers to use is complicated by the number of ways in which film reading can be undertaken (see box).

No randomized controlled trials have been conducted to examine this question specifically; a systematic review of 10 cohort studies showed that double reading increased the rate of cancer detection (Dinnes *et al.*, 2001), but cost–effectiveness remains an open issue (see Chapters 2 and 6). The policies of national screening programmes vary, but two readers are used in most. For example, in The Netherlands, all films are read independently at a central unit by two radiologists. Consensus between the two readers is required for a referral, which may contribute towards the low referral rate in that country.

Number of readings and readers

Technique	No. of readers
Single reading	2
Double reading with recall if necessary	2
Double reading by censensus	2
Double reading with arbitration	3

In the United Kingdom, the policy is for single reading, although, in practice, some kind of double reading is used in most programmes. Radiographers (technicians trained to take radiographs) may also be trained as film readers, particularly where too few radiologists are available. While issues such as screening interval, number of mammographic views and the lower age limit for screening have been further evaluated in specific trials, the issue of double reading has not, and it is likely to remain an open question. Studies will probably be undertaken to evaluate film reading by non-radiologists and with use of computer-aided detection (see Chapter 2).

Clinical breast examination

In most population-based screening programmes, mammography is the only method used for detection, although clinical breast examination is added in some countries or regions (see Chapter 2 and Shapiro *et al.*, 1998). No randomized trials of clinical breast examination versus no screening have been completed (see Chapter 4).

Some form of clinical breast examination was used in the Health Insurance Plan, Edinburgh and Canadian trials. In the Health Insurance Plan study, 67% of cases were detected by clinical examination (with or without mammography), but in the UK Trial of Early Detection of Breast Cancer, which included the intervention arm of the Edinburgh trial, the relative sensitivity of clinical breast examination was only 70% at first screens and 44% at subsequent screens (Moss *et al.*, 1993). As shown in Chapter 4, no difference in the reduction in mortality was found in trials with clinical breast examination and those with mammography alone.

The Canadian trial of women aged 50–59 was specifically designed to compare annual mammography plus clinical examination with clinical examination only (Miller *et al.*, 1992b). No difference was found in the rate of death from breast cancer between the two arms of the trial after 13 years of follow-up, although more small tumours without node involvement were detected with mammography, and the confidence interval of the relative risk for breast cancer mortality was wide (RR, 1.02; 95% CI, 0.78–1.33) (Miller *et al.*, 2000).

Clinical examination is used alone in Japan, where annual screening is carried out by medical practitioners. Recent evidence suggested that this can reduce mortality rates (Kuroishi *et al.*, 2000). Screening by clinical breast examination alone has also been proposed for developing countries with limited resources (Mittra *et al.*, 2000).

Breast self-examination

There is no evidence from randomized trials that breast self-examination is effective in reducing breast cancer mortality (see Chapter 4). A study in Shanghai, China, in which women employed in factories were assigned randomly to a self-examination instruction group or to a control group on the basis of factory, showed no difference in cumulative breast cancer mortality at 10–11 years (Thomas *et al.*, 2002). Nevertheless, a number of countries or programmes include it in their recommendations.

Indicators for monitoring and evaluating the effectiveness of screening programmes

An organized screening programme should have access to an information system that covers the programme and the entire target population. The background measures of coverage and attendance relate to the target populations and the women in it, whereas performance indicators, such as predictive value and detection rate, are related directly to the mammography unit and other diagnostic facilities. Mortality from

Why the reduction in mortality due to screening takes longer to become evident in national mortality statistics than in randomized controlled trials

Reason for delay in mortality reduction from national screening programme compared with trials	Comments and comparison with trials
Dilution due to breast cancer deaths in cases diagnosed before any invitation to screening	Pre-exisiting cases are excluded from both arms of trials.
Long time to cover national population, e.g., in the United Kingdom, the first invitations were sent between 1988 (start of programme) and 1995 (completion of prevalence round), depending on area of residence.	Women enter trials at time of first invitation, which is time zero.
Learning time for many staff new to screening	Trials usually have highly experienced staff.

breast cancer, excess mortality and total mortality are the indicators that cover the entire process. As they are related directly to the purpose of screening, mortality is the necessary and sufficient indicator of effectiveness. All other measures are necessary, and may be early indicators that the programme is operating as expected, but they are not sufficient and cannot as such be taken as proof of effectiveness.

The aim of any breast cancer screening programme is to reduce breast cancer mortality. However, such a reduction will take many years to emerge in a population-based screening programme, starting from a few years after introduction of the programme but taking decades to show a full effect. The effect is much slower than in randomized controlled trials, for the reasons shown in the box below. Implementation of national screening programmes has tended to be slow: the United Kingdom and The Netherlands started screening in 1988 and 1990, respectively, but all women were not screened until 1995 and 1997.

The reduction in breast cancer mortality rates due to screening at a given time after the start of the programme is complex to measure because, in the absence of screening, breast cancer mortality in a defined age group is affected by a number of factors. These include cohort effects, improvements in treatment, presentation at an earlier stage as a direct result of the introduction of the programme and the attendant publicity (Stockton *et al*., 1997) and changes in death certificate coding. Furthermore, the effect of screening in reducing breast cancer mortality, as seen in national statistics, will be diluted by deaths among women in whom breast cancer is diagnosed before an invitation to screening. If record linkage to a cancer registry is available, 'refined breast cancer mortality' can be used, which excludes deaths among women in whom cancer was diagnosed before the start of screening (Hakama *et al*., 1999). A further complexity is that, although the reduction in breast cancer mortality seen in trials is related to the age of the women at entry into the trial, national breast cancer mortality statistics are a measure of the decrease in breast cancer mortality rates of women at the age at which they die. Therefore, some effect is seen in increasingly older women with time since the start of screening. Useful information is derived by comparing breast cancer mortality among women invited to screening with that of women who were not invited and that among invited women who participate with that of non-participants. However, these comparisons can be biased by differential access to treatment by uninvited women and by a differential cancer risk of non-attenders, unless proper controls are found.

Therefore, early indicators are needed to ascertain whether adjustments are required to a screening programme in the early stages. These indicators of performance can be used to predict the final reduction in breast cancer mortality that is likely to be achieved with the current level of screening performance. National screening programmes in countries with relatively small populations, and therefore large statistical uncertainty in breast cancer mortality rates by 5-year age band, and in which no control group is included are unlikely to allow a reliable estimate of the effect of screening unless indicators of performance are used.

Origins of indicators of effectiveness of screening

Nearly all measures of the performance of breast cancer screening programmes are compared with target or expected values, which are derived, either explicitly or implicitly, from information from randomized controlled trials. Use of the results of such trials is essential, as they have produced well-defined reductions in breast cancer mortality. Application of the parameters of these trials, adjusted for local conditions, allows prediction of the eventual reduction in breast cancer mortality likely to be achieved in a national programme. In general, the targets or standards for measures such as cancer and interval cancer detection rates in a particular population-based screening programme should take into account age range, background incidence and screening interval. Accordingly, separate calculations are required for each population-based screening programme.

Some of the first suggestions were made by Day *et al.* (1989) on the basis of experience from the Swedish Two-county trial. In the latest follow-up, there was a 32% (95% CI, 20–41%) reduction in breast cancer mortality in the invited compared with the control group (Tabár *et al.*, 2000b). Table 42 lists these indicators of performance in chronological order of availability of data. Proactive evaluation of breast cancer screening requires evaluation of data on an annual basis, even after the first year of screening, to determine whether corrective action is required. The European guidelines for performance are given as examples in Tables 43 and 44.

Table 42. Measures of monitoring

Measure	Type of evaluation provided
Participation (or compliance) rate	Indicates potential for effectiveness of the overall programme
Prevalence rate at initial screening test and rate of interval cancers	Provides estimates of sensitivity, lead time, sojourn time and predictive value
Stage (or size) distribution of screen-detected cancers	Indicates potential for reduction in absolute rate of advanced cancers
Rate of advanced cancers	Early surrogate of mortality
Breast cancer death rate	Final evaluation

From Day *et al.* (1989)

Table 43. Indicators for assessing the performance of a breast cancer screening programme for women aged 50–64

Performance indicator	Acceptable level	Desirable level
Participation rate	> 70%	> 75%
Technical repeat rate	< 3%	< 1%
Recall rate		
Initial screening	< 7%	< 5%
Subsequent regular screening	< 5%	< 3%
Additional imaging rate at time of screening	< 5%	< 1%
Pre-treatment diagnosis of malignant lesions	> 70%	> 90%
Image-guided fine-needle aspiration cytology procedures with insufficient result	< 25%	< 15%
Benign:malignant biopsy ratio		
Initial screening	≤ 1:1	≤ 1:1
Subsequent regular screening	≤ 0.5:1	≤ 0.2:1
Re-invitation within the specified screening interval	> 95%	100%

From Commission of the European Communities (2001)

Performance indicators

Participation

The first important indicator of performance is attendance to screening (also called participation, compliance or uptake). Determination of this indicator can be deceptively complicated. For example, in a programme in which women aged 50–69 are screened every 2 years, 10 possible screening invitations can lead to 210, or 1024, different possible attendance patterns, all attending and all not attending being the extremes. If a woman attends only two screens at the ages of 50 and 52, the potential benefit will be clearly different from that of a woman who attends at the ages of 66 and 68. This is true for all variations in attendance pattern. The indicators used currently

Table 44. Early surrogate indicators for assessing the effectiveness of a breast cancer screening programme for women aged 50–64

Surrogate indicator	Acceptable level	Desirable level
Interval cancer rate/background incidence		
0–11 months	30%	< 30%
12–23 months	50%	< 50%
Breast cancer detection rate		
Initial screening	3 x incidence rate	> 3 x incidence rate
Subsequent regular screening	1.5 x incidence rate	> 1.5 x incidence rate
Stage ≥ II/total cancers detected at screening		
Initial screening	25%	< 25%
Subsequent regular screening	20%	< 20%
Invasive cancers ≤ 10 mm/total invasive cancers detected at screening		
Initial screening	≥ 20%	≥ 25%
Subsequent regular screening	≥ 25%	≥ 30%
Invasive cancers/total cancers detected at screening	90%	80–90%
Node-negative cancers/total cancers detected at screening		
Initial screening	70%	> 70%
Subsequent regular screening	75%	> 75%

From Commission of the European Communities (2001)

tend to be simplified measures of attendance, although they are likely to be reasonably accurate within the context of their use.

When participation in screening is used to predict reduction in breast cancer mortality, the participation rate is compared with the target measure. For example, in the United Kingdom National Health Service Breast Screening Programme, the percentage of invited women who attended for screening was about 70% (Blanks *et al.*, 2002), while in the Swedish Two-county trial it was about 90%; the relative attendance is therefore 0.78. If equivalent rates of detection of invasive cancer, or interval cancers, are assumed after allowing for differences in background incidence between the United Kingdom and Sweden, then the estimated reduction in the United Kingdom would simply be 0.78 x 30% = 23%. This calculation is based on the assumption that there is no major effect of selection bias. Calculation of selection bias requires information on breast cancer incidence rates among non-attenders.

Estimated reduction in breast cancer mortality based on cancer detection and participation rates

Cancer detection rates are the first indicator of screening performance and, if monitored and evaluated on an annual basis, give the earliest information on achievable mortality reduction. They are subject to overdiagnosis bias (see later in this chapter), but participation and cancer detection rates can give the earliest indication of possible under-performance of a regional or national programme. Corrective action can then be taken early, rather than waiting for other indicators, such as interval cancer rates, which can be measured only several years later. It is important to distinguish invasive from non-invasive cancers, as a high rate of detection of invasive cancers is the principal measure of interest. Cancer detection rate targets should take into account age range, background incidence and screening interval, and separate targets are required for prevalence and incidence screens.

Detailed evaluations were made of the screening programme in the United Kingdom in 1995 by using the background rates in England and Wales to estimate the expected rates of detection of invasive cancers and interval cancers (Blanks *et al.*, 1996; Moss & Blanks, 1998). The standardized detection ratio was introduced, in which indirect age standardization was used to calculate the expected number of invasive cancers that would indicate parity with the Swedish Two-county study. The ratio was used to evaluate the performance of regional screening programmes among the 95 programmes in the United Kingdom, after adjustment of the expected number of invasive cancers by local background incidence (Blanks & Moss, 1996). In practice, the standardized detection ratio was found to be a good quality assurance measure for detecting under-performing programmes. A standardized detection ratio of < 0.75

(the lowest recorded statistically stable ratio being 0.5) was taken to indicate possible under-performance and the need for a visit by quality assurance staff. On its own, a low standardized detection ratio merely suggests under-performance and might be misleading if screened women have a different distribution of risk from the target population. In the United Kingdom, this appears rarely to be the case, but it might hold in other countries.

Table 45 shows the observed and expected numbers of invasive cancers in the United Kingdom annually between 1 April 1993 and 31 March 2000 for women aged 50–64 (Blanks *et al.,* 2000a). The participation rate in the United Kingdom was about 70%, and the final reduction in mortality was estimated from the relative uptake x standardized detection ratio x 30%. In 1993–94, this was (70/90) x 0.83 x 30% = 19.4%; by 1999–2000, it had risen to (70/90) x 1.14 x 30% = 26.6%. Previous work showed that, over the range 0.8–1.3, the standardized detection ratio and interval cancer rates are strongly correlated (Given-Wilson *et al*., 1999), which justifies these calculations. The reduction in mortality is that which would have been achieved in a comparable randomized controlled trial, given similar diagnosis and treatment.

Table 45 shows that, in the United Kingdom, the rate of detection of invasive cancers was inadequate in the early years of the programme, as confirmed by the high rates of interval cancers. This low rate was not observed earlier partly because of a high rate of detection of DCIS, which contributed to achievement of what was considered to be an adequate cancer detection rate. In fact, the high DCIS detection rate 'masked' a poor rate of detection of small invasive cancers. This example illustrates why the rates of detection of invasive cancers and small invasive cancers should be considered separately (Day *et al.,* 1995) and cautiously because of overdiagnosis. Overdiagnosis is commonly associated with DCIS but is also likely to occur in the case of some invasive cancers and strongly correlates with stage of disease. A further measure used in the United Kingdom is the standardized detection ratio (< 15 mm), which is the ratio of the observed number of invasive cancers < 15 mm to that expected from the Swedish Two-county study. Alternatively, a standardized detection ratio for invasive cancers < 10 mm can be used.

Table 45. Observed and expected numbers of invasive cancers in women aged 50–64 in United Kingdom National Health Service breast screening programme

Screening year	Observed	Expected	SDR	Modelled mortality reduction (%)[a]
1993–94	4447	5344.6	0.83	19.4
1994–95	4452	4952.5	0.90	21.0
1995–96	4486	4725.5	0.95	22.2
1996–97	4833	4799.7	1.01	23.6
1997–98	5187	4964.2	1.04	24.3
1998–99	5744	5064.2	1.13	26.4
1999–2000	5795	5076.6	1.14	26.6

From Blanks *et al.* (2000a); data for 1999–2000 are unpublished
SDR, standardized detection ratio
a Estimated on the basis of 70% participation

Estimated reduction in mortality based on interval cancer rates and participation

Interval cancers are those which present in the interval between screens after a negative screen. The rate can be expressed either as that of interval cancers or as a proportion of the expected incidence rate (had screening not been undertaken). These estimates assume the existence of cancer registration. Poor quality cancer registration and/or record linkage can lead to underestimation of interval cancer rates and therefore overestimation of programme performance. If it is assumed that the data are of sufficient quality, the expected reduction in mortality can be calculated from participation and the combined proportionate incidence of interval cancers. If the proportionate incidence is x% in the first year, y% in the second and z% in the third, the combined proportionate incidence is (x + y + z)/3.

Table 46 shows data for the Anglia region of the United Kingdom on interval cancer rates during the early years of screening (Day *et al.,* 1995). The background incidence in the absence of screening was estimated as 2.2 per 1000. The combined proportionate incidence was (0.24 + 0.59 + 0.79)/3 = 0.54, indicating that 54% of the incidence expected in the absence of screening was observed. The Dutch screening programme, with a 2-year interval, gave similar estimates for the early years of screening, with proportional incidences of 27% and 52% in the first and second years (Fracheboud *et al.*, 1999). The combined proportionate incidence can be used to estimate the expected reduction in mortality in conjunction with the participation rate.

Day *et al*. (1995) were then able to calculate the expected reduction in mortality in the early years of screening

Table 46. Interval cancer rates during early years of screening in the United Kingdom (Anglia region)

	Time since last negative screen (months)		
	0–11	12–23	24–35
Interval cancer rate per 10 000 women–years	5.2	12.8	18.9
Proportionate incidence	0.24	0.59	0.79

From Day *et al.* (1995)

by inference to the Swedish Two-county study, as shown in Table 47. The participation rate in Anglia was 80%, which gave an estimated reduction in mortality of 21% at a combined proportionate incidence of 54%. As 70% participation is achieved in the United Kingdom as a whole, an 18% reduction can be expected, with a combined proportionate incidence of 54%. This reduction is closely in line with the estimate of 19% from screening in 1993–94 by use of the standardized detection ratio and participation rate (see Table 45). This is to be expected, as the standardized detection ratio and interval cancer rates are highly inversely correlated (Given-Wilson *et al.,* 1999).

Estimated reduction in breast cancer mortality on the basis of prognostic factors

Reduction in incidence of advanced cancer

If screening programmes are successful in allowing earlier diagnosis of cancer, an overall reduction in the rates of advanced cancer should be observed in the target population. This should result in reduced mortality from advanced disease. Day *et al.* (1989) suggested that differences in stage distribution by mode of detection would appear immediately, an effect on advanced cancer rates at diagnosis would appear only about 4 years after initiation of screening, and an effect on mortality some 2 years later, i.e. 6–7 years after the onset of screening. On the basis of observations in the Swedish Two-county trial, they suggested that screening should decrease the rate of advanced (stages II–IV) tumours by at least 30% after 4 years.

Obtaining information on stage

Stage has major prognostic implications. It is based on several factors: size of the tumour mass, its degree of spread both locally and to distant (metastatic) sites and involvement of regional lymph nodes; it is recorded according to the TNM system (UICC, 2002), American Joint Committee on Cancer (2002) stage (I–IV) or the summary 'extent of disease' (local, regional, distant). In the last scheme, tumours classified as TNM T2N0M0 or stage IIA of the American Joint Committee on Cancer (2 < 5 cm, localized to the breast) are considered 'localized'. Thus, in different studies, advanced disease may correspond to stage II–IV or IIB–IV.

Cancer registries do not always include information on stage of cancer. Registries may vary in the quality of information on stage and its completeness (proportion of 'unstaged' cases) (Berrino *et al.*, 1995) and over time. This must be taken into account in comparative studies of incidence by stage, including time trends. Comparisons over time may be also be biased by the increasing availability of techniques for staging, so that cancers that might have been described as localized with less sophisticated diagnostic techniques are now described as advanced. This phenomenon is known as 'stage migration'. As the size of a primary tumour is less subject to this type of bias, it is the measure preferred by many workers for evaluating stage, with corresponding prognostic implications for the patient. Size is ideally assessed from resected pathological specimens, as described in Chapter 1.

Rates, not proportions

Comparisons of prognostic factors (size, stage) should be presented as rates per population screened, as opposed to percentages. Rates allow consideration of changes in the proportions of cancers detected by screening. In the early phase of a screening programme, when most examinations are prevalence screens, a high proportion of small or early-stage cancers will be detected (and, in consequence, a decreased percentage of advanced cancers). Similarly, significant 'overdiagnosis'

Table 47. Expected reduction in mortality from screening in the United Kingdom (Anglia region) on the basis of participation and interval cancer rates

Interval cancers (combined proportionate incidence)	Mortality reduction with 70% participation	Mortality reduction with 80% participation
0.34	24	29
0.40	22	26
0.50	19	22
0.54	18	21
0.60	15	18
0.66	13	15

From Day *et al.* (1995)

of small lesions would lead to a decreased percentage of advanced tumours, altough the absolute rates may be unchanged. Expression of results as the percentage reduction in the incidence of advanced cancers requires an estimate of the incidence of advanced cancers that would have been observed in the absence of the screening programme.

Time trends in the incidence of breast cancer by stage of disease

In an analysis of data from the Surveillance, Epidemiology and End Results (SEER) programme in the USA for 1973–93, significant decreases in mortality were seen after 1989 in all age groups (a slight increase in rates had preceded this). The possibility that screening may have been partly responsible was suggested by the increased incidence of localized disease and a subsequent decline in the incidence of regional disease in women in each age group over 40. By 1990, more than 40% of women had received a mammogram in the previous year, mainly for screening purposes (Chu *et al.*, 1996).

In Limburg, The Netherlands, the annual number of breast cancers diagnosed increased by almost 50% immediately after the introduction of screening and then decreased to previous levels after completion of the first screening round. There was a 10% decrease in the incidence of stage II–IV tumours and a 15% decrease in tumours with node involvement over those seen in the period directly before screening began (1987–90). The incidence of tumours with node involvement was 1% lower in 1994 and 15% lower in 1995 (Schouten *et al.*, 1998).

In the study in East Anglia, United Kingdom, the increase in the incidence of small cancers in the early years of screening was much greater than the subsequent decrease in the incidence of advanced cancer, suggesting that the reduction in mortality might have been somewhat lower than that targeted (McCann *et al.,* 1998; Table 48). The authors used three methods to estimate the 'expected' incidence of advanced cancer in the absence of screening. The first was a projection from the incidence observed in 1976–86 to that in 1995. For the second, they took the average rate of advanced cancers observed in 1987–88, immediately before the onset of screening, and compared it with the 1995 rate. Finally, they took the ratio of advanced to early cancers in 1989–94 in women who had not yet received an i nvitation to screening, generated an expected incidence rate of advanced cancers, and multiplied this by the actual number of cases in order to obtain the number of cases expected. The predicted rates of advanced cancers from data for 1987 and 1988 suggested a reduction of about 20%, while the predicted rate with exclusion of 1987 and 1988 indicated a much smaller reduction (5.3%).

The screening programme in New South Wales (Australia) was started in 1989. Between 1984 and the end of 1995, an estimated 72% of women in their 50s and 67% of women in their 60s had had at least one mammogram in the organized screening programme or in the private health system (Kricker *et al.,* 1999). Before 1989, the incidence of breast cancer increased only slightly (+1.3% annually), but between 1990 and 1995 it increased more rapidly (+3.1% annually). Between 1986 and 1995, the rates of small cancers (< 1 cm) increased steeply, by 2.7 times in women aged 40–49 and by 5.6 times in women aged 50–69. The incidence of large breast cancers (≥ 3 cm) up to 1995, after little apparent change up to 1992, fell by 17% in women aged 40–49 and by 20% in those aged 50–69 years. Mortality from breast cancer increased slightly between 1972 and 1989 (+0.5% annually) but then fell (–2.3% annually) between 1990 and 1995 (Kricker *et al.,* 1999). The decline in the incidence of advanced cancer was not, however, seen overall in 1995–97 (Coates *et al.*, 1999).

For countries without organized screening programmes, monitoring and evaluation have certain requirements and limitations. The minimum level of information needed to make some

Table 48. Incidence rates per 100 000 of invasive breast cancer by TNM stage in women aged 50–69 years, East Anglia region, United Kingdom

Stage	Year of diagnosis								
	1981–86	1987–88	1989	1990	1991	1992	1993	1994	1995
Early stage (stage I)	55.5	76.1	89.1	141	160	142	123	141	119
Advanced (stages II, III, IV)	131	162	159	140	129	110	134	135	132
Total (including unknown stage)	196	245	256	289	297	257	263	282	253
Proportion of advanced (%)	70	68	64	50	45	44	48	49	53

From McCann *et al.* (1998)

estimate of performance is an indication of routine attendance for mammography (every 3 years or less), obtained from sample surveys with questionnaires, if no other source of information is available. Cancer registration provides some indication of screening activity. Trends in rates of advanced disease, as shown in the SEER programme in the USA (Chu *et al.*, 1996) are informative.

The screening programme in Tuscany, Italy (1970–97), showed a small decline in the rate of advanced cancers, but the timing and the fact that it occurred throughout the Province and not just where the organized screening programme had been introduced, suggested that the changes were the result of widespread 'spontaneous' early detection activities (Barchielli & Paci, 2001).

Estimation of reduction in mortality from observed distributions of tumour grade, size and nodal status (surrogate measures)

With the availability of detailed information on tumour size, grade and node status, a more sophisticated estimate can be made of the reduction in mortality. The technique requires, however, an uninvited comparison group as well as detailed information on survival in relation to size, grade and nodal status. This technique was used by McCann *et al.* (2001) with results from the Anglia region of the United Kingdom, where the introduction of screening was staggered by district and by year of birth. There were thus sufficient numbers of women in the region and in the age group targeted for screening who did not receive a first invitation until well after the start of screening in the region in 1989. The technique is more complex than estimates based on interval cancer rates and detection rates. The results suggested that screening in Anglia would reduce mortality by around 7% in women aged 50–54 at diagnosis and by 19% in those aged 55–64 at diagnosis. Overall, for women aged 50–64 at diagnosis, the reduction would be 15%. However, the technique is sensitive to the lead time for the screen-detected cases: using a 3-year lead time rather than 2 years gave an overall reduction in mortality of 19% in women aged 50–64 at diagnosis. The method is also dependent on assumptions about long-term survival, improved diagnostic classification (stage migration) and confounding by treatment.

Table 49 summarizes the estimates of mortality reduction with the three techniques described above. For Anglia, the techniques result in a range of estimates, from 15% to 21%. In the United Kingdom as a whole, the two simple estimates are 19% and 18%. The broad conclusions are similar in all cases: that the programme in the United Kingdom was less effective in the early years of screening than in the Swedish Two-county study. The performance of the United Kingdom programme has increased markedly in recent years for a number of reasons. First, use of two views and double reading has increased; secondly, standardization of film density has improved image quality; thirdly, under-performing programmes have been identified with the standardized detection ratio method and have been improved; fourthly, radiologists have become more experienced at both film reading and assessment during the decade since the start of the programme (Tabár *et al.*, 1995; Blanks & Moss, 1999).

Table 49. Estimated final reduction in mortality from breast cancer for women aged 50–64 at diagnosis, with three techniques, from the early screening data in the United Kingdom

Technique (reference)	Population	Mortality reduction (%)
Standardized detection ratio plus participation (Blanks *et al.*, 2000a)	United Kingdom	19
Interval cancer rate plus participation (Day *et al.*, 1995)	Anglia	21
Interval cancer rates plus participation (Day *et al.*, 1995)	United Kingdom	18
Grade, size and nodal status (2-year adjusted lead time) (McCann *et al.*, 2001)	Anglia	19
Grade, size and nodal status (3-year adjusted lead time) (McCann *et al.*, 2001)	Anglia	15
MISCAN model[a] (van den Akker-van Marle *et al.*, 1999)	United Kingdom Netherlands	24 29

From Tabár *et al.* (1995); Blanks & Moss (1999)

[a] The estimates of mortality reduction from the MISCAN model are based on the estimated sensitivity of the screening test and the screening interval and age range of the invited women. The estimate for the United Kingdom is higher than that with the other techniques because of the poor sensitivity of the screening programme in the early years.

Any of these methods could be used to estimate mortality reduction, and, ideally, all the methods would be used sequentially as the data became available. The standardized detection ratio method is useful for timely estimates of screening performance and can result in rapid implementation of quality assurance checks if the ratio is too low. It has the further advantage that the calculations are very simple once the target rates for each age range and type of screen have been calculated. Use of interval cancer rates as a performance indicator takes longer but is in theory a better method, provided the data are of sufficient quality. Finally, surrogate measures represent the most sophisticated method for estimating the likely reduction in breast cancer mortality that will be achieved.

MISCAN model

In The Netherlands, the effectiveness of screening has often been estimated with the MISCAN approach (van den Akker-van Marle *et al.,* 1999), in which a simulated population is used which represents the demographic characteristics and the breast cancer incidence and mortality of the population under study. The natural history of breast cancer is modelled as a progression through successive disease states. Indicators of screening programme performance (e.g. attendance and sensitivity) are added to the model, and the effects on breast cancer mortality with and without screening are estimated. It has been estimated with the MISCAN model that breast cancer mortality in The Netherlands would decrease in women aged 55–74 by 5% in 1996, by 18% in 1999 and by 29% in the long term.

Many of the above conclusions were reached by comparing performance with that in the Swedish randomized trials (as the gold standard) and modelling experience with intermediate indicators to the expected mortality reduction (de Koning *et al*., 1995b).

Mortality from and screening for breast cancer in different countries

Screening for breast cancer with mammography is based on the evidence from several randomized trials (see Chapter 4) that showed reductions in mortality from breast cancer of about 20–30% for women aged ≥ 50. The important question is whether these results are reproducible as a public health policy, by applying mammography in routine screening. It is likely that the organization, the quality of the technique and the devotion and skills of the persons running a routine programme are different from those in a scientific trial.

Routine screening programmes can be evaluated most readily by time trends and differential mortality from the disease for which screening is being performed. Probably the best known is screening for cervical cancer. The substantial differences among the Nordic countries in the extent of organized screening were closely matched by the mortality rates from cervical cancer (Läärä *et al.*, 1987).

Screening for breast cancer is done either as an organized public health policy or by more spontaneous activity. The International Breast Cancer Screening Network (Shapiro *et al.*, 1998; Klabunde *et al.*, 2001b) comprises 22 countries with national, regional or pilot programmes for screening with mammography. The best known are those in the Nordic countries, The Netherlands and the United Kingdom. Regional efforts are being made in Italy and other southern European countries, and spontaneous activity is widespread, e.g. in Germany and the USA. The implementation period of these programmes is described in Chapter 3.

One of the first papers to report the correlation between routine screening and breast cancer mortality was that of Quinn and Allen (1995) in England and Wales. A change in the trend of mortality was found at the time screening was introduced, whereas there should have been a lag between screening and death if the prolongation of life was due to screening. Hence, the change in trend was too early and probably due to a change in the national treatment policy. Up to the late 1990s, the trend in mortality from breast cancer was linear, with no major indication of an effect of mammography (Figure 31). However, in a detailed analysis, Blanks *et al.* (2000b) estimated a 6% reduction in breast cancer mortality due to screening among women aged 55–69 in 1998, in a programme which covered the population between 1988 and 1995. This estimate is likely to represent the beginning of the effect of screening in the United Kingdom.

In Sweden, the screening programme started as a cluster-randomized trial in two counties (Tabár *et al.*, 1985); individually randomized trials in various parts of Sweden soon followed (Nyström *et al*., 1993). A study based on geographical differences in mortality rates in Sweden and a comparison of the results of the original trials showed an estimated reduction in breast cancer mortality of 19%, i.e. somewhat less than those reported in the original trials (Törnberg *et al*., 1994).

Jonsson *et al*. (2001) compared counties in Sweden and estimated a 20% reduction in breast cancer mortality due to screening in women aged ≥ 50. The estimate was based on deaths only among women in whom breast cancer was diagnosed after the start of the programme. This refined mortality rate is not readily available in routine statistics and, furthermore, assumes the availability of a cancer registry and data protection legislation that allow linkage of the two data sources. As only 27% of all deaths from breast cancer occurred among women in whom breast cancer was diagnosed after the start of the programme, the effect on overall breast cancer mortality can be estimated to be 5–6%, which is clearly too small an effect to be readily identifiable in routine statistics (Figure 32). It is close to the 6% arrived at in the United Kingdom by Blanks *et al*. (2000b). Sjonell and Stahle (1999) found no reduction in mortality in Swedish national data. Given the smaller population and therefore greater instability of breast cancer mortality rates in Sweden, an early effect of breast cancer screening is unlikely to be seen.

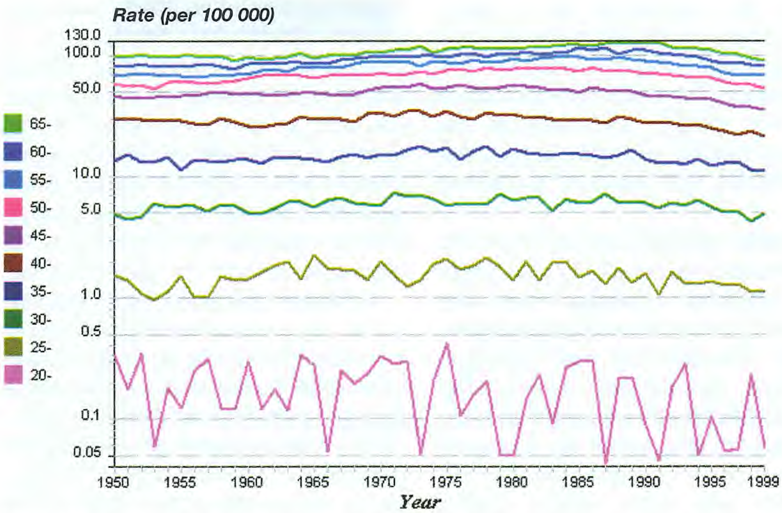

Figure 31 Trends in breast cancer mortality by age, United Kingdom, 1950–99
From WHO (1999b)

The effect of routine screening on mortality from breast cancer was also difficult to estimate in The Netherlands (van den Akker-van Marle *et al*., 1999). In a recent study, Broeders *et al*. (2001) evaluated the effect of the screening programme that started in Nijmegen a quarter of a century ago. Mortality from breast cancer between 1969 and 1997 was analysed and compared with data for Arnhem, with no such programme, and for The Netherlands as a whole. Data were not available on deaths among patients in whom breast cancer was diagnosed before screening started, but the long follow-up and use of elegant modelling techniques reduced the bias due to these data. Depending on the comparison group (Arnhem or The Netherlands) and the assumptions used in the model, a nonsignificant 6–16% reduction in mortality was estimated.

The acceptability of using a whole country as the control implies that the effect of the national programme on the risk for death from breast cancer is small; this is confirmed in Figure 33. In 2001, the National Evaluation Team for Breast Cancer Screening reported that the first significant reduction in breast cancer mortality had occurred in 1997–99, among women aged 55–74, of 7–13% in comparison with the pre-screening period of 1986–88. This effect was, however, less than the estimate arrived at with the MISCAN model (Fracheboud *et al*., 2001a).

As mentioned earlier, a substantial fall in mortality from breast cancer was seen in Australia (Kricker *et al*., 1999), which was not totally attributable to screening. The difficulty of evaluating effectiveness can be exemplified by the situation in Finland, where overall breast cancer mortality can be specified for those populations first subjected to screening, then by mortality refined for prescreening diagnosis and finally for those women first invited to screening, with preselected (i.e. unbiased) controls at each stage. A nationwide population-based screening programme for breast cancer was started in 1987 and gradually

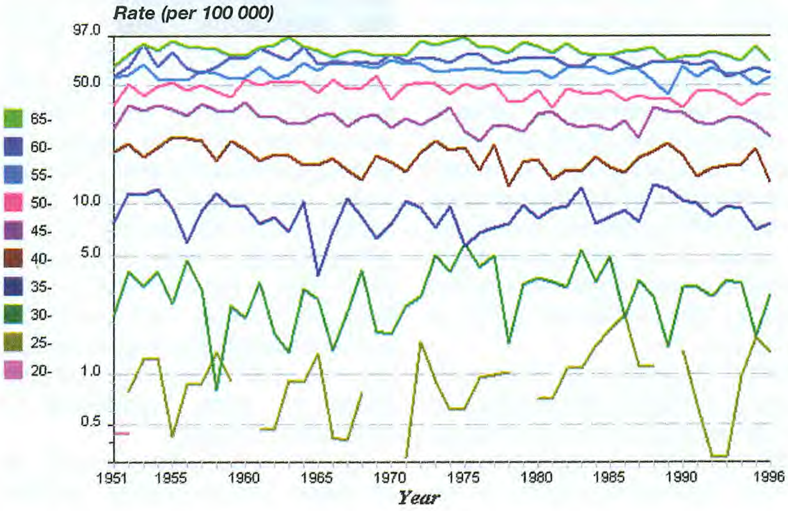

Figure 32 Trends in breast cancer mortality by age, Sweden, 1951–96
From WHO (1999b)

extended to cover all women aged 50–64 (Hakama *et al.*, 1991). Women in 1-year birth cohorts recommended by the National Board of Health were identified individually and invited for screening, and the same women were re-screened every 2 years. In 1987, it was recommended that women born in 1928, 1932 and 1936 should be screened, and the programme was expanded to cover all the other even-year birth cohorts. More age cohorts were included in the programme during the implementation phase.

A centralized, comprehensive information system is provided by the Mass Screening Registry within the nationwide Finnish Cancer Registry for identification, invitation and follow-up of women and for evaluation of the effectiveness of the programme. Cancer registration is virtually complete (Teppo *et al.*, 1994). The National Population Registry, the National Register of Deaths and the Cancer Registry are linked with the screening results via the Mass Screening Registry. Intermediate indicators derived from this programme, such as attendance, specificity and sensitivity, show good quality (Pamilo *et al.*, 1990; Saarenmaa *et al.*, 1999).

National figures are appropriate for evaluating the Finnish programme because the policy is nationwide and the programme was implemented for a relatively short time. No obvious change in national trends in mortality from breast cancer corresponding to the screening programme was seen (Figure 34) in the crude data, but a more refined analysis is needed.

Any change in mortality rates should first be seen in women born in 1928, 1932 and 1936—that is, in the cohorts that were screened first, in 1987, the first year of the public health policy. As screening was delayed for several years and for a minimum of 2 years among women born in the adjacent cohorts of odd birth years, they were selected as controls. It was assumed that any effect would be seen some years after screening but before the controls benefitted from screening. As the mortality rates by birth cohort were similar (Figure 35), the cohort-specific deaths did not indicate any effect of screening.

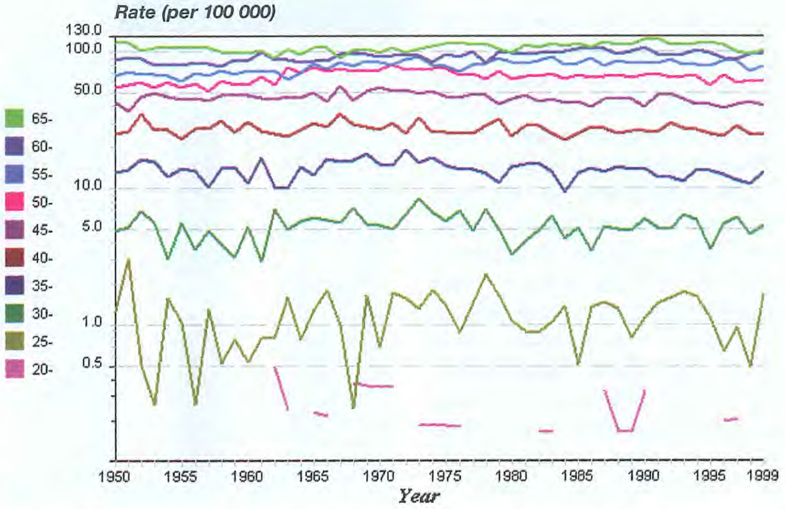

Figure 33 Trends in breast cancer mortality by age, The Netherlands, 1950–99
From WHO (1999b)

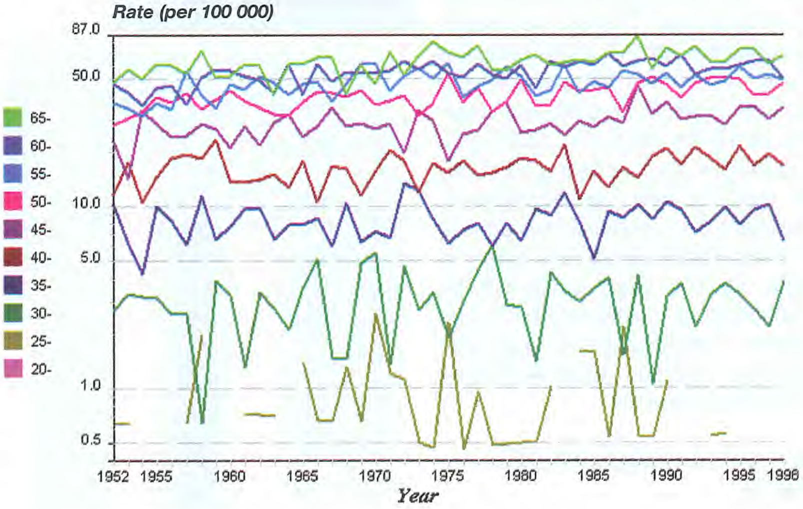

Figure 34 Trends in breast cancer mortality by age, Finland, 1952–98
From WHO (1999b)

Screening will affect only deaths from breast cancer among women in whom breast cancer was diagnosed after the start of screening. There was no substantial difference in mortality between the target and the control populations when incident cases diagnosed before 1987 were excluded (refined mortality rate) (Figure 36).

The design of the Finnish programme allowed identification of individual women by screening status and by date of invitation to screening (Hakama *et al.*, 1997). The participants in the programme were women born between 1927 and 1939. Invitations to screening were given in 1987–89, and the invited women were classified as screened or non-responders. The controls were women in the same municipalities, matched for age with those screened and individually identified at the same time as women invited for screening. By the end of 1992, 64 women invited for screening and 63 controls had died from breast cancer diagnosed after the start of follow-up. The refined breast cancer mortality rate was lower for those invited to screening than for the controls (RR = 0.76), indicating a 24% protective effect of screening. This was not statistically significant (95% CI, 0.53–1.09). For the cohorts born in 1932 and after, the effect was larger (RR = 0.56) and statistically significant (95% CI, 0.33–0.95). The protective effect appeared relatively early, from the third year of follow-up. More details of this trial-like evaluation of the public health policy are given in Chapter 4. The effect could not be demonstrated by routine statistics owing to dilution and selection. A study with randomized, matched individual controls was therefore essential.

Alternative measures of effect on mortality

A number of alternative measures derived from the estimated reduction in mortality due to screening may be useful in decision-making (Hakama *et al.*, 1999). These were calculated for the Finnish population, in which mammographic screening has been estimated to have achieved a 24% reduction in breast cancer mortality among women aged 50–64 years who were invited to

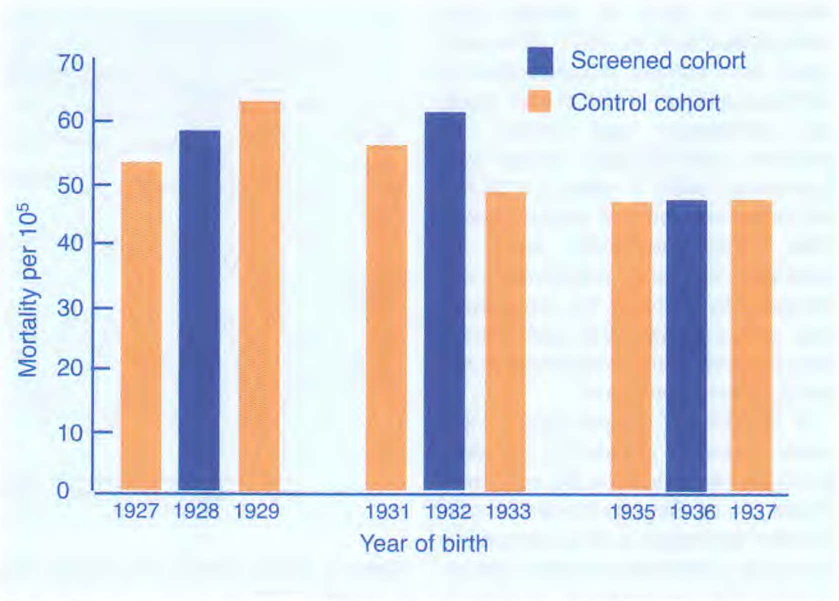

Figure 35 Birth cohort-specific mortality rates from breast cancer (per 100 000 woman–years) in Finland, 1990–95
From Hakama *et al.* (1999)

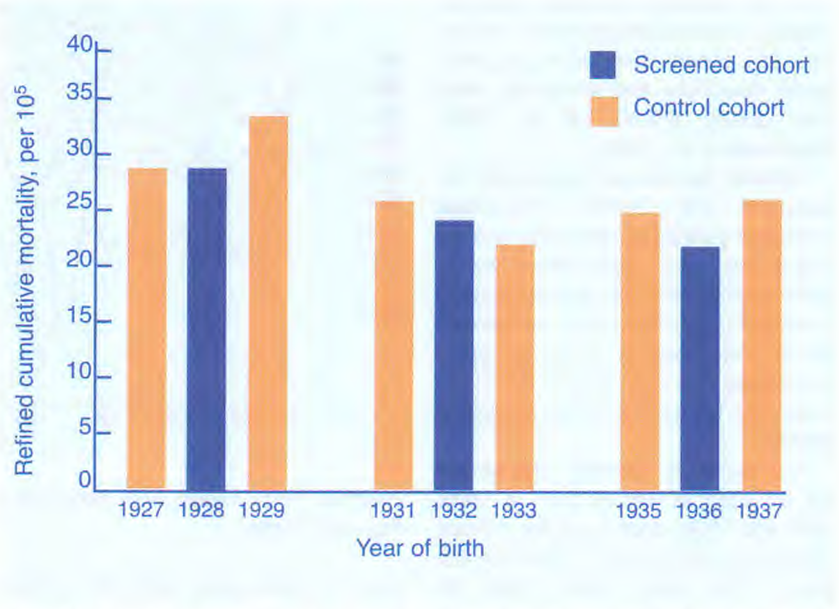

Figure 36 Cumulative (refined) mortality rates from breast cancer (per 100 000 woman–years) in Finland by birth cohort during 1987–95 after exclusion of incidence cases diagnosed before 1987.
From Hakama *et al.* (1999)

screening (Hakama *et al.,* 1997). These measures include: the number of cancer deaths prevented per screen (estimated to be four deaths per 10 000 screens), life span gained per breast cancer death prevented (estimated to be 15 years), life span gained per patient with breast cancer detected by screening (estimated to be 1.5 years); life span gained per screen (estimated to be 2.2 days, which can be compared with the estimated half day spent by a woman attending for screening) and life span gained per invitation to screening, i.e. per member of the target population of women (estimated to be 1.9 days).

Balance between false-positive and false-negative results

Screening quality must be evaluated in parallel with estimates of breast cancer mortality reduction. While the emphasis in the United Kingdom screening programme tended to be on detection rates of invasive cancers, the screening programme in The Netherlands focused on prevention of too many false-positive results. Verbeek *et al*. (1991) suggested that the "first measure to pay attention to is the specificity. If the specificity does not meet the reference value, improvements have to be made irrespective of the other control outcomes [positive predictive value and detection rate]. In such a screening set up [a high] proportion of healthy women with a positive screening test is not acceptable." They therefore introduced simultaneous evaluation of performance and quality.

The screening programme in The Netherlands started in 1988, around the same time as that in the United Kingdom, and is notable for its very low recall rate for assessment (de Koning *et al*., 1995b). Verbeek *et al.* (1991) suggested a target recall rate of < 1%, while the Forrest (1986) report indicated that the acceptable recall rate in the United Kingdom could be as high as 10%. In the United Kingdom, variations in positive predictive value, referral rates for diagnostic confirmation (also called recall rate) and cancer detection rates from individual programmes were studied with charts showing positive predictive value of referral against referral rate, with the cancer detection rate expressed as isobars (Blanks *et al*., 2001). The variation in individual programmes for both measures was shown to be very high, the positive predictive value ranging from 26% to 6% and the recall rate from 2% to 9%. Programmes tended to have similar results each year. The diagrams suggest that, in the United Kingdom, a positive predictive value of 25% is too high and results in a marginally lower detection rate. Programmes with recall rates of about 4%, positive predictive values of 15–20% and standardized detection ratios of around 1.3 achieved the highest quality of screening. It is clear that some individual programmes have better quality screening. In many cases, it is possible to suggest how detection rates could be improved. Like the standardized detection ratio, the positive predictive value–referral diagram is useful as an internal quality control measure for centres with similar risk distributions in their target populations. In the United Kingdom, the emphasis has been on detection rates and then on improving screening quality once those rates have been achieved. This was true particularly after the high interval cancer rates reported in the early years of screening.

It is interesting to compare the screening programmes in The Netherlands and the United Kingdom in terms of quality. It should be noted that the procedures used in the two countries are different, and some caution should be exercised in comparing the percentages of women referred and positive predictive value. Nevertheless, in The Netherlands, there has been a strong effort to maintain high specificity (Verbeek *et al*., 1991) and, as a consequence, a high positive predictive value and a low referral rate. The low referral rates are based partly on the low rates in the pilot study and the Nijmegen study and show the influence of those studies on national screening programmes.

Table 50 shows the results of The Netherlands screening programme between 1990 and 1995. The referral rates are very low indeed, particularly at subsequent screens, and this is acknowledged as a feature of the national programme. However, the cancer detection rate at subsequent screens was 20% lower than expected, and the low referral rates at subsequent

Table 50. Comparison of observed and expected results of The Netherlands screening programme, 1990–95, for women aged 50–69

	Initial screens		Subsequent screens	
	Observed	Expected	Observed	Expected
Referrals (% of screened women)	1.4	1.6	0.7	0.6
Positive predictive value of referral (%)	48	41	51	57
Detection rate per 1000 women screened	6.6	6.5	3.4	4.3

From Fracheboud *et al.* (1998). Expected values estimated from MISCAN model screening rounds have been postulated as a possible explanation (Fracheboud *et al.*, 1998).

The screening programme in the United Kingdom provides an interesting contrast. Table 51 shows the data on subsequent screening in The Netherlands and the equivalent data for the United Kingdom in 1994–95. The comparison is interesting, although it also illustrates that great care should be taken in in interpreting performance data among countries, as women are referred differently in The Netherlands and the United Kingdom. Nevertheless, the dramatic difference is surprising, as is the difference in cancer detection rates. Further work is required to enable a more useful comparison. It is debatable whether screening performance measures, particularly those related to quality, can be adequately compared across countries.

Factors such as screening quality vary not only among countries but also dramatically within countries. In the United Kingdom, referral rates can vary by 2–9% and positive predictive value by 6–26% (Blanks *et al.*, 2001). Giordano *et al.* (1996) reported data derived by applying the performance measures in the European guidelines (Commission of the European Communities, 2001) to the Italian breast screening programmes and found that most have attained 'acceptable' or 'desirable' levels for many indicators. The differences in indicators of screening quality show that the interpretation of the measures is to some extent subjective. How many false-positive referrals should be tolerated for each cancer detected? In The Netherlands, the number is very low, but in the United Kingdom it depends on the clinician in charge of the screening centre. Setting screening quality targets is much more subjective than setting targets for screening effectiveness in reducing breast cancer mortality. Furthermore, measurement of screening quality is complicated by differences in the screening protocols among countries. Comparison of referral rates is complex, even when the national organized programmes are superficially not very different, as in The Netherlands and the United Kingdom. Any comparison with, for example, the USA, which does not have a national organized screening programme, would therefore be invidious.

In contrast, measures of screening effectiveness in terms of reducing mortality, e.g. participation, cancer detection rates (standardized detection ratios), interval cancer rates (as a proportional incidence measure) and stage distribution of screen-detected cancers, can be compared across countries more readily.

Table 51. Observed and expected values for subsequent screens: Netherlands (1990–95, age 50–69) and United Kingdom (1994–95, age 50–64)

	United Kingdom		Netherlands	
	Observed	Expected[a]	Observed	Expected[b]
Referrals (% of screened women)	3.4	≤ 7	0.7	0.6
Positive predictive value of referral (%)	13	–	51	57
Detection rate per 1000	4.3	> 3.5	3.4	4.3

From Fracheboud *et al.* (1998); Blanks *et al.* (2000a)
[a] In early years
[b] Based on MISCAN model

Conclusion

Screening programmes should ultimately be monitored in terms of deaths, the measure directly related to the purpose of screening. The effect of screening is real but small at present, the estimates of change in national overall breast cancer mortality rates being 5–10% in countries with the longest experience. The estimates were larger in a few studies of sub-populations and after removal of bias due to deaths in cases diagnosed before the start of screening. The gain in life years per screen is nevertheless likely to remain small. Small reductions in breast cancer mortality, usually < 10%, will increase with length of follow-up and may ultimately approach the estimates found in randomized trials, of 20–30%. As such results will take a long time to achieve, the change will be very gradual and probably not immediately visible in national trends. Prolongation of follow-up will not affect the small estimated time gained in comparison with time spent, as screening is usually repeated every 2 years.

Although screening for breast cancer may thus appear to be insufficiently effective for use as a public health policy, that conclusion is probably not justified. Screening for breast cancer also has a humanitarian value, in addition to the prolongation of life. Screening, in principle, offers a greater chance to select the type of intervention, including breast-conserving and less invasive treatment. Most recalls are due to false-positive results, which cause unnecessary anxiety and invasive or otherwise unpleasant investigations. A decision on whether to screen should depend on a weighting of all the effects and how they compare with other health services. Many other health activities have not been properly evaluated and may be even less effective.

Reliable monitoring of a screening programme should be based on death as the outcome indicator and on measures derived from deaths. When such an approach is not possible, surrogate outcome indicators should be used, although favourable results based on established standards do not necessarily imply a reduction in breast cancer mortality. Only in special circumstances will it be possible to distinguish the component of the reduction in mortality that can be attributed to screening from other effects, such as treatment. Screening with mammography prevents some deaths from breast cancer. The effect is certain but small. In terms of prolongation of life, the effect is about 2 days per woman per screen. In terms of standardized mortality ratios, the effect may approach that seen in trials and ultimately a reduction in breast cancer mortality of about 20%.

Hazards (risks) of screening

The underlying rationale for breast cancer screening is to promote health by identifying women with breast cancer at an early enough stage that treatment will cure the disease. However, the vast majority of women undergoing screening do not have breast cancer at the time of the examination, and these women cannot derive a direct health benefit from screening; they can only be harmed. The following sections address the two major categories of possible harm that are relevant to any programme of early detection: false-positive results and overdiagnosis. In addition, although a diagnosis of breast cancer earlier than its clinical presentation is part of the pathway to potential benefit, it also implies that women have to live longer knowing that they have a potentially serious disease. For some women, this is balanced by more conserving surgery and improved survival or cure, but for the majority is represents only a disadvantage. The effect of an earlier diagnosis of disease on the quality of life is an immediate negative aspect of screening, against which any prolongation of life should be weighed (Figure 37).

Two possible harms specific to mammography are also considered: an early increase in mortality from breast cancer and radiation-induced cancer.

Occurrence and consequences of false-positive results in mammography

False-positives and overdiagnosis

The term 'false-positive' refers to an abnormal mammogram (one requiring further assessment) in a woman ultimately found to have no evidence of cancer. 'Overdiagnosis' refers to the diagnosis and treatment of cancers that would never have caused symptoms. Thus, a false-positive result can be found only in a woman without cancer, while overdiagnosis can be made only for women with cancer. While an individual woman can readily be identified as having had a false-positive result, overdiagnosis can never be identified reliably for an individual, as virtually all abnormalities labelled 'cancer' are treated.

The 'harm' of false-positive mammograms relates to the additional testing, invasive procedures and anxiety that would never have happened in the absence of screening. The 'harm' of overdiagnosis relates to unnecessary anxiety (associated with a diagnosis of potentially fatal disease) and unnecessary treatment. Both harms are inevitable if a screening programme is to be effective (Morrison, 1992). The challenge is to minimize both while still detecting those cancers for which early diagnosis and treatment can alter the clinical course of disease.

Definition of a false-positive rate

Two definitions have been used to define a false-positive screening result. The broad definition includes all mammograms that are accompanied by a recommendation for further assessment (e.g. repeat examination, clinical exami-

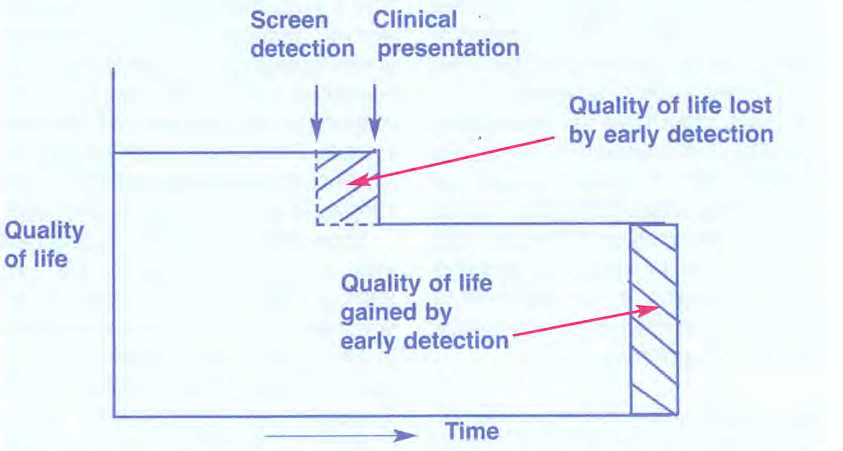
Screen detection
Clinical presentation
Quality of life lost by early detection
Quality of life
Quality of life gained by early detection
Time

Figure 37 Quality of life over time for women whose breast cancers are found clinically and by screening

nation, diagnostic mammogram, ultrasound or breast biopsy) for women who are found ultimately not to have cancer. A narrower approach is to count only those recommendations that ultimately lead to a breast biopsy in these women (Kopans, 1992). While breast biopsies are arguably the most important adverse effect of a false-positive mammogram, the broader definition is used here, both because it is commoner and because it more completely captures all subsequent events.

The frequency of false-positive mammograms is conceptually measured only among women ultimately found not to have cancer and is defined as:

$$\frac{\text{Number of exams with a recommendation for further assessment among women known not to have cancer}}{\text{Number of exams among women known not to have cancer}}$$

In practical terms, however, the same number is well estimated as a product:

$$\text{\% of all mammograms requiring further assessment} \times (1 - \text{PPV})$$

where PPV is the positive predictive value or the percentage of mammograms requiring further assessment that are ultimately found to be cancer.

Either approach can be interpreted as answering the question: 'What is the probability that a healthy woman will require further assessment after a single screening mammogram?' Although this measure is, strictly speaking, a probability or a proportion, in the remainder of this chapter it is referred to by its familiar label: the 'false-positive rate'.

False-positive rates observed in practice

The reported false-positive rates range from < 1% to > 10% (Table 52). Two broad observations can be made from these data. First, false-positives are more frequent at a woman's first screening examination than at subsequent examinations. This observation probably reflects the value of having a previous film for comparison and, in national programmes, cumulative experience in mammography.

Secondly, different groups of radiographers perform differently. In particular, false-positive rates are an order of magnitude higher in community practice in the USA than in The Netherlands screening programme. This difference may reflect different thresholds for recommending further evaluation of specific mammographic abnormalities, such as calcifications and well-circumscribed nodules. For example, the false-positive rate will increase if further evaluation is suggested for smaller, less characteristic clusters of calcifications. These differences may, in turn, be explained by the distinct medico-legal climate in the USA, where a missed diagnosis of breast cancer is now the commonest and second most costly basis for malpractice suits (Black *et al*., 1995; Mitnick *et al.,* 1995; Physician Insurers Association of America, 1995).

Cumulative risk

Most false-positive results are reported from a single mammogram. However, as women undergo screening mammograms repeatedly, an individual woman's cumulative risk of ever having a false-positive results increases with repeated screens. From the woman's perspective, therefore, it may be important to know the cumulative risk for a false-positive result.

Some data are available. During the 4 years of the Health Insurance Plan programme, about 5% of women in the screened group had a recommendation for biopsy after a false-positive mammogram (Shapiro *et al*., 1988a). In the Stockholm trial, approximately 1% of biopsies conducted as a result of false-positive mammograms were performed in women invited for two annual rounds of screening (Lidbrink *et al*., 1996). This low rate was seen after only two screens, but there was also a low rate of abnormal mammogram readings (0.8–1.8%) in the Stockholm study. In the Screening Mammography Program of British Columbia, Canada, the cumulative risk for a false-positive mammogram after 10 screens was estimated to be 38% for women aged 40–49, 35% for women aged 50–59 and 29% for those aged 60–69 (Olivotto *et al*., 1998).

In the USA, Elmore *et al.* (1998) studied the experience of 2400 women screened in a health plan in Massachusetts. After a 10-year follow-up, 23.8% of the women had had at least one false-positive result and 5.1% had had an invasive procedure as a result of a false-positive result. Using a Bayesian version of a product or an estimate of the Kaplan-Meier type, in which mammographic screening events were used instead of time, the authors estimated that, after 10 mammograms, 49% (95% CI, 40–64%) of the women would have had a false-positive result. When the definition of a false-positive result was limited to women without cancer who underwent a breast biopsy, the cumulative risk over 10 mammograms was estimated to be 19% (95% CI, 10–41%).

The same general approach was used to estimate the cumulative risk of ever having a false-positive result as a function of the two most relevant inputs: the false-positive rate (in which first and subsequent screens are distinguished) and the number of times screening is repeated (a function of the screening interval; Table 53). The effects of various conditions on the cumulative risk for a false-positive result are clear.

Adverse consequences of false-positive results

The adverse effects reported after false-positive results include increased use of health care and increased patient anxiety.

Table 52. Chance that a women without breast cancer will require further assessment after a single mammogram

Setting	Proportion of		False-positive rate (% abnormal x (1–PPV)) (%)
	Mammograms requiring further assessment (% abnormal)	Abnormal examinations in which cancer is diagnosed (PPV) (%)	
National programmes			
Netherlands (de Koning *et al.*, 1995b)			
First screen	1.4	48–51	0.7
Subsequent screens	0.9	36–54	0.5
United Kingdom (Blanks *et al.*, 2000a)			
First screen	7–8	6–8	7.0
Subsequent screens	3–4	12–14	3.0
US National Breast and Cervical Cancer Detection Program (May *et al.*, 1998)			
First screen	5	9.5	4.5
Subsequent screens	4	5.6	3.8
Other			
US academic practice (Kerlikowske *et al.*, 1993)			
First screen (age 50–59)	7	9	6.4
First screen (age 60–69)	8	17	6.6
Subsequent screens (age 50–59)	2	16	1.7
Subsequent screens (age 60–69)	2	7	1.9
US Medicare (age 65–69) (Welch & Fisher, 1998)			
Mixture of first and subsequent screens	8.5	8	7.8
US community practice (Brown *et al.*, 1995)			
Mixture of first and subsequent screens	11	3.5	10.6

PPV, positive predictive value

Increased cost and health-care use

False-positive results are associated with increased numbers of office visits, diagnostic mammograms, ultrasounds and breast biopsies. Lidbrink *et al.* (1996) reported that 502 women with false-positive results in the Stockholm trial made 1539 visits to a physician and had 542 fine-needle aspiration biopsies, 257 diagnostic mammograms and 118 surgical biopsies. The cost of evaluating the false-positive results was 26.5% that of screening. In the study of Elmore and colleagues (1998), 631 false-positive results generated 601 office visits, 384 diagnostic mammograms, 176 breast ultrasounds, 100 open or core biopsies, 28 fine-needle aspirations and one hospitalization. In the same study, it was estimated that about US$ 33 would be spent on follow-up procedures to evaluate false-positive results for every US$ 100 spent on screening mammography. May *et al.* (1998) found that abnormal mammogram results generated additional mammographic views in 56%, sonography in 31%, clinical examinations in 30%, fine-needle aspirations in 8.7% and breast biopsy in 28% of cases. These percentages were not broken down according to true- and false-positive results, and no costs were included.

None of these studies included the costs of increased health-care use by patients. Barton and colleagues (2001) found that, in the 12 months after recommended follow-up, false-positive results were associated with more patient-initiated visits for both breast-related (incidence ratio, 4.03; 95% CI, 2.97–5.47) and non-breast-related (incidence ratio, 1.18; 95% CI, 1.09–1.28) reasons, including mental health services. Extrapolating to women

Table 53. Estimated cumulative risk of ever having a false-positive result on mammography under various conditions

False-positive results (%)				
First screen	Subsequent screens	20-year programme of screening with examination every:		
		3 years	2 years	1 year
0.7	0.5	3.2	5	10
2.0	1.0	7	10	19
4.0	2.0	13	20	35
6.0	3.0	19	29	47
8.0	4.0	25	36	58
10.0	5.0	30	43	66
10.6	10.6	49	67	89

The cumulative risk is calculated as 1 minus the chance of never having a false-positive result (which, in turn, is the product of the probabilities of having a normal result in multiple examinations). For example, consider the upper left-hand cell — the probability of ever having a false-positive result of a women screened every 3 years for 20 years in a programme with a false-positive rate of 0.7% on the first screen and 0.5% on subsequent screens. The chance of not having a false-positive result is 99.3% on the first examation and 99.5% on each subsequent screen. The cumulative risk over 20 years in which six examinations are performed is 1 – (0.993 x 0.995 x 0.995 x 0.995 x 0.995 x 0.995) or 3.2%.

eligible for screening in the population of the USA, the authors estimated that false-positive results could generate as many as 14.4 million non-breast-related physician visits over a decade.

Emotional and psychological effects

Rimer and Bluman (1997) reviewed nine studies conducted before 1997 which specifically addressed the psychological impact of false-positive results of mammography. Using the same search strategy, the Working Group identified nine more studies (Table 54). The studies vary by country, the type of patients studied, when in relation to the mammogram they were studied and the instruments used to determine their psychological state. Nonetheless, all but one showed transient negative psychological effects associated with a false-positive result.

Most of the studies showed that the increase in anxiety was moderate. Although the increase was short-lived in most women (e.g. Brett *et al.*, 1998; Gilbert *et al.*, 1998; Olsson *et al.*, 1999), some experienced longer-term consequences of a false-positive result. In one study, an increased anxiety score was reported 3 months after an abnormal mammogram, and in another the score was still increased 18 months after the test (Gram *et al.,* 1990; Lerman *et al.,* 1991).

Few studies have addressed the impact of a false-positive result on behavioural measures. One study showed that women who were recalled were more likely to continue practising breast self-examination (Bull & Campbell, 1991). Barton *et al.* (2001) reported that physicians were more likely to record breast-related concern for women who had had a false-positive result and that these women were more likely to use health-care services, for both breast and non-breast-related problems.

Three studies have been conducted of the effect of a false-positive result on future screening behaviour. Burman and colleagues (1999) compared the subsequent adherence to screening mammography of 813 women who had had false-positive results and 4246 women who had had normal mammograms. After adjustment for multiple risk factors, the women who had had a false-positive result were slightly more likely to attend for their next screening mammogram than women with a normal result (odds ratio, 1.21; 95% CI, 1.01–1.45). Pisano *et al.* (1998) surveyed 43 women who had undergone excisional breast biopsy after receiving a false-positive result 3 years earlier. When compared with 53 randomly selected women with normal mammograms and 83 women with 6-month recall, the women who had had a biopsy were slightly more likely to attend for subsequent screening mammography. Lerman *et al.* (1991) also found that more women who had had a false-positive result than those with normal screens attended for their next scheduled screening mammogram (74–78% versus 68%, p > 0.05). No studies were found of screening behaviour after repeated false-positive results.

Clearly, substantial proportions of women who have a false-positive result become anxious about breast cancer. This was true in several countries and cultures. Anxiety tends to be greatest at the time of notification of an abnormality and less (or resolved) when the work-up is completed without breast cancer being found. There is no evidence that false-positive results decrease future adherence to screening recommendations and in fact may increase it slightly. Women may therefore understand that false-positive results are a part of mammography. Schwartz *et al.* (2000) found that 99% of 479 women were

Table 54. Studies of psychological status and health behaviour after a false-positive result on a screening

Reference and country	Type and time of measurement	Groups and numbers of women and response		Response rate (%)	Results			
						% anxious:		
						Normal	False-positive result	
Ellmann *et al.* (1989)	General health questionnaire in	Normal mammogram:	295	Overall, 98	At visit	35	44	
United Kingdom	person	False-positive result:	271		3 months	26	29	
	At visit and 3 months later	Breast cancer:	134		later			
					$p < 0.02$–0.02 at visit			
					% anxious about		% practising	
					breast cancer		self-examination	
Bull & Campbell	Mailed self-administered question-	Invited to screening:	750	72	5		10	
(1991)	naire	Normal mammogram:	420	79	4		10	
United Kingdom	Before screening	False-positive result, no			2		24	
	6 weeks after screening	invasive test:	240	72	6		35	
		False-positive result, biopsy:		68	Not significant		$p < 0.00$	
					Retrospecitive anxiety score:			
Sutton *et al.* (1995)	Mailed general health questionnaire				False-positive		Normal	
United Kingdom	with 7-item anxiety subscale				result			
	Before screening	1021		68	1.6		1.6	
	At screening	795		78	1.7		1.3	
	9 months after screening	795		78	1.1		1.1	
Swanson *et al.* (1996)	Psychological consequences					Mean psychological score		
United Kingdom	questionnaire, mailed and on-site					Invited	Screened	Recalled
	At invitation	False-positive result:	33	49	Physical	0.7	0.2	3.0
	At screening			68	Emotional	1.3	0.5	4.1
	At recall			100	Somatic	1.1	0.5	3.3
					* All recall scores significantly differ from earlier scores			
					% psychological consequences			
Ong *et al.* (1997)	Mailed psychological	Regular recall:	130	Overall, 75	29			
United Kingdom	conseqences questionnaire	after assessment:	128		50			
	1 month after final visit	after fine-needle						
		aspiration:	106		58	$p < 0.0005$		
		Early recall after						
		assessment:	130		63			
		Regular recall after biopsy:	30		87			

Table 54 (contd)

Reference and country	Type and time of measurement	Groups and numbers of women and response		Response rate (%)	Results	
					% psychological consequences	
Brett *et al.* (1998) United Kingdom	Mailed psychological consequences questionnaire 5 months after screening	Normal mammogram	52	Overall, 76%	10	
		False-positive result, non-non-invasive work-up:	51		45	
		False-positive result, fine-needle aspirate cytology:	41		44	$p < 0.0001$
		False-positive result, biopsy:	45		59	
		False-positive result, 6-month recall:	23		61	
Gilbert *et al.* (1998) United Kindom	Mailed and (on-site hospital) anxiety and depression scale and health questionnaire				% anxious:	
	Before screening	2110		90	39	
	At screening	1463		70	31	
	At recall	122		98	47	
	5 weeks after recall	90		74	34	
	4 months after recall	90		74	33	
					$p < 0.02$–0.001 for recall vs others	
					% anxious about breast cancer:	
Gram *et al.* (1990) Norway	Mailed self-administered questionnaire 18 months after screening	Normal mammogram:	152	73	13	
		False-positive result:	126	79	29	
					$p = 0.001$	
					% anxious about breast cancer	
Gram & Slenker (1992) Norway	Mailed self-administered questionnaire 1 year after screening	Normal mammogram:	209	84	22	
		False-positive result:	160	89	40	
					$p < 0.05$	
Lidbrink *et al.* (1995) Sweden	Self-administered questionnaire: and blood tests on site				Mood score:	
	At follow-up test				Time !: 2.3	
	3 weeks after completed work-up	False-positive result:	48	98	Time 2: 3.4	
					$p < 0.05$	
					No difference in cortisol or prolactic concentration or in lymphycytic stimulation	

Table 54 (contd)

Reference and country	Type and time of measurement	Groups and numbers of women and response		Response rate (%)	Results		
					Mean psychological score		
Olssen *et al.* (1999) Sweden	Mailed psychological consequences questionnaire 1 and 6 months after final visit	False-positive result:	235	93		1 month	6 month
		Normal mammogram:	987	89		~ 0.75	~ 0.3
						~ 0.19	~0.17
					$p < 0.001$ between false-positive result and normal mammogram		
Lerman *et al.* (1991) USA	Anxiety questionnaire by telephone interview Mammography adherence by telephone interview				% anxious about mammography	% attendance at next screen	
	3 months after false-positive result:	Normal mammogram:	121	Not reported	48	68	
	All work-up completed	False-positive result, low suspicion:	119		61	78	
	15 months later	False-positive result, high suspicion:	68		70	74	
					$p = 0.008$	$p < 0.05$	
Pisano *et al.* (1998) USA	Record of intention for screening mammogram, by telephone interview 3–4 years after index screening				% attended screen x 3 years	% intended regular future screens	
		False-positive result and excisional biopsy:	43	Overall, 75	72	98	
		False-positive result and 6-month recall:	83		58	82	
		Normal mammogram:	53		66	90	
					$p = 0.26$	$p = 0.036$	
					% returned for recommended screen:		
Burman *et al.* (1999) USA	Screening mammogram, by computerized record review Up to 6 months after next recommended screening	False-positive result:	813	Overall 85	73		
		Normal mammogram:	4246		74		
					% medical records with anxiety about breast cancer noted		
Barton *et al.* (2001) USA	Office visits by medical record reviewer For 12 months after screening	Normal mammogram:	496	100 (medical records)	0.2		
		False-positive result:	496		10.0		
					$p = 0.001$		
Cockburn *et al.* (1994) USA	Mailed and on-site psychological consequences questionnaire At screening Before results At recall clinic 1 week after 'all clear' 8 months later	False-positive result:	58	70	Mean emotional and physical dysfunction scores significantly increased in group with false-positive results at recall and 1 week after 'all-clear'. Otherwise, scores in 3 groups similar		
		Normal mammogram:	142	68			
		'Community':	52	72			

Table 54 (contd)

Reference and country	Type and time of measurement	Groups and numbers of women and response		Response rate (%)	Results
Lowe *et al.* (1999) Australia	Mailed, self-administered questionnaire, general health questionnaire, psychological consequences questionnaire				Women with false-positive results more concerned about breast cancer and more so than for normal mammo gram after 1 month ($p < 0.05$)
	At appointment	3158	95		
	At recall	False-positive result:	182	94	
		Normal mammogram:	182	81	
	1 month after recall	False-positive result:	182	81	
		Normal mammogram:	182		
Scaf-Klomp *et al.* (1997) Netherlands	Psychological consequences questionnaire for false-positive result Self-administered questionnaire for normal mammogram				Women with false-positive result scored higher on psychological consequences questionnaire than those with normal mammogram but no higher than randomly selected women without mammogram
	8–10 weeks after first mammogram	False-positive result	74	78	
	6 months after first mammogram	Normal mammogram:	113	59	
		Randomly selected women without mammogram:	238	59	

aware that false-positive results occur, and most accepted them as a consequence of screening. This was true regardless of whether the respondent had actually experienced a false-positive result.

Decreasing the adverse effects of false-positive results

As false-positive results cannot be totally eliminated, strategies to reduce the occurrence and severity of the adverse psychological effects and behaviour should be developed. Lindfors *et al.* (2001) compared stress in women who had undergone immediate work-up and received a false-positive result and women who returned for later work-ups and reports. The mean overall stress rating on a five-point scale was 2.3 for women who had undergone immediate work-up and 2.8 for women who returned later for work-up. The response rate to the survey was 40%, but the two groups of women who did respond were similar in terms of demographic variables. Women receiving immediate follow-up evaluation after an abnormal result may not have time to become anxious before the result is clarified, and some may be unaware that an extra view or ultrasound is conducted because an abnormality was found on their screening mammogram. Immediate follow-up requires the presence of a radiographer, and the cost implications of this strategy should be assessed if the finding that immediate follow-up reduces anxiety is repeated.

Ong and Austoker (1997) studied the effect of discussion of results with nurses when patients were recalled for further evaluation after an abnormal mammogram. Fewer women who had a chance to talk with a nurse wanted to talk later about why the assessment was needed (4%) than women who did not talk to a nurse (30%), and fewer wanted additional information. The result was similar when an information leaflet was added to the recall letter (Austoker & Ong, 1994).

When information was given in the leaflet, women perceived less need for more information. Whether educating women about false-positive results in general or making counselling services available prevents or lowers anxiety is not known and should be studied.

Decreasing the rate of false-positive results

Although the literature on strategies for lowering the false-positive rate is sparse, and no formal trials of strategies were found, it would appear reasonable to address the risk factors for false-positive results. The literature suggests that characteristics relating to both the subject and the mammography process are involved.

Factors related to women that affect the false-positive rate

Age is inversely related to the false-positive rate (Kerlikowske *et al.*, 1993; Kopans *et al.*, 1996; Lidbrink *et al.,* 1996), at least partly because dense breasts are more difficult to read radiographically (Fajardo *et al.*, 1988). However, it was found in one study (van Gils *et al.*, 1998) that no such difficulty was apparent for mammograms read after 1983, when compared with those read between 1975 and 1982. No stratification by age was reported, which could be important, because average breast density decreases slightly as women age and epithelial tissue is replaced with fatty tissue (Tabár & Dean, 1982). White *et al.* (1998) demonstrated that the breast density of premenopausal women was greater during the luteal (2 weeks before onset of menses) than the follicular (2 weeks after onset of menses) phase of menses.

Several studies have shown that postmenopausal hormonal replacement therapy also increases breast density (Berkowitz *et al.*, 1990; Stomper *et al.,* 1990; Kaufman *et al.*, 1991; McNicholas *et al.*, 1994; Laya *et al.*, 1995; Greendale *et al.*, 1999). Four studies documented an increased frequency of false-positive readings in postmenopausal women on hormone replacement therapy. Laya and colleagues (1996) found a relative risk for a false-positive mammogram reading of 1.71 (95% CI, 1.37–2.14) in current users and 1.16 (95% CI, 0.93–1.45) in former users, versus never users of estrogen replacement therapy. Christiansen *et al.* (2000) found similar effects, while Kavanagh *et al.* (2000) found that, in comparison with non-users, users of hormone replacement therapy had an adjusted odds ratio of 1.12 (95% CI, 1.05–1.19) for a false-positive result. Thurfjell *et al*. 1997 found decreased mammographic specificity, especially in women treated with both estrogen and progesterone replacement.

Having had a breast biopsy was associated with a higher risk for a false-positive result (Brenner & Pfaff, 1996), although one prospective study of recall rates among women with and without a history of breast biopsy showed no difference (Slanetz *et al.*, 1998). Christiansen *et al.* (2000) found a risk for a false-positive result of 20% for women with a history of three or more breast biopsies, 13% for women with two, 11% for women with one and 6.1% for women with no history of a breast biopsy ($p < 0.01$).

Factors related to mammography that affect the false-positive rate

Perhaps the most important effects on the risk for a false-positive result are related to the diagnostic ability of the radiographer. Brown *et al.* (1995) found that the frequency of reading screening mammograms in 50 individual practices in a representative national sample in the USA ranged from a low of 3% to a high of 57%. Using a model with adjustment for multiple variables relating to the woman, Christiansen *et al*. (2000) studied 35 community radiologists and estimated that the odds of a woman having a false-positive result was 11-fold higher (95% CI, 2–17) if the film was read by the radiologist with the highest false-positive rate than when it was read by the radiologist with the lowest rate. While some of this variation undoubtedly reflects the small sample studied (mean of 35 films per radiologist in both studies), the variation is nonetheless substantial.

The availability of previous mammogram films for comparison when reading mammograms has been shown to decrease the frequency of false-positive results. Christiansen and colleagues (2000) found that the false-positive rate was halved when previous films were available. Frankel *et al.* (1995) found that the frequency of abnormal results dropped from 7% at initial examination to 3% on subsequent examinations at which previous films were available.

Only one report (Christiansen *et al.*, 2000) was available of the combined effect of patients' risk profile and radiological variables on a woman's risk for a false-positive result in multiple screens. The characteristics examined included the age, history of previous breast biopsy, family history of breast cancer, menopausal status, estrogen use, body mass index, race and median household income of the patient and comparison with previous mammogram, time since last mammogram and radiologist's recall rate. In a multivariable model, four patient variables emerged as independent risk factors. The risk decreased with the patient's age and increased with the number of breast biopsies, family history of breast cancer and estrogen use. In addition, all three radiological variables were independent risk factors. A woman with average risk factors had a 15% chance of having a false-positive result by the ninth mammogram if her films were read by a radiologist with a low recall rate, and an 86% chance if her mammograms were read by a radiologist with a high recall rate. This study was of one setting and may not be generalizable to others. Also, breast density was not included as a variable. Although many of these features are immutable, all three radiological variables associated with false-positive readings are modifiable.

Banks *et al*. (2002b) predicted the recall rate after a false-positive result for 60 000 women in the United Kingdom who were not using hormone replacement therapy. Premenopausal and perimenopausal women were more likely than postmenopausal women to be recalled for false-positive results, and the variation in recall rate by age was accounted for by the menopausal status of the women. Furthermore, women were more likely to be recalled if they had had breast surgery in the past, and less likely to be recalled if a comparison mammogram was available. There were also weak associations with parity and weight, but other factors, including educational level, family history of breast cancer, tobacco and alcohol consumption, height, age at birth of first child, breastfeeding history and past use of hormonal contraceptives had no effect on the recall rate.

Possible strategies for decreasing false-positive rates

It may be reasonable to alter estrogen use in the short term, thereby decreasing the false-positive rates for women on hormonal replacement therapy. In a preliminary study, Harvey *et al*. (1997) found that stopping hormone replacement therapy for 10–30 days before a repeat mammogram resulted in resolution of or a decrease in mammographic abnormalities in 35 of 47 patients.

Changing mammographic practices in settings with high false-positive rates is probably a more relevant option. One way of lowering the frequency of false-positive results is to set explicit goals for lowering recall rates. In the USA, the Agency for Health Care Policy and Research has recommended that the recall rate be no more than 10% of screening mammograms (Bassett *et al*., 1994a). In Europe, the Europe Against Cancer Programme suggested that the 'acceptable' level of recall after the first screen should be < 7%, and the 'desirable' level should be < 5% (Commission of the European Communities, 1996). Lowering recall rates may be easier in Europe than in North America because of the tendency for malpractice suits in the latter.

It has been suggested that more experienced radiologists with a higher volume of mammographs have lower false-positive rates (Sickles *et al*., 1990), but this requires confirmation. There is evidence that the availability of previous mammograms and interpretation of the image by two radiographers (see above) can decrease the false-positive rate. Both of these ideas should be weighed against increased costs, the feasibility of obtaining previous mammograms (Bassett *et al*., 1994b) and the costs of more professional input.

Overdiagnosis

An obvious source of harm associated with any screening programme is unnecessary treatment of cancers that were not destined to cause death or symptoms. This section describes the concept of overdiagnosis and reviews the evidence for overdiagnosis of breast cancer after screening mammography. The section concludes with a description of autopsy series in which there was a reservoir of undetected breast cancers, which might be diagnosed as imaging techniques become capable of detecting progressively smaller lesions. The following section describes mammographic detection of DCIS, for which overdiagnosis may be particularly common.

The concept of overdiagnosis

Overdiagnosis refers to the detection of cancers that would never have been found were it not for the screening test (Prorok *et al*., 1999). Patients in whom such indolent cancers are detected do not benefit from screening and can only experience harm: the worry associated with a 'cancer' diagnosis and the complications of therapy. For most prospective screenees (and many clinicians), overdiagnosis is a foreign concept. This is understandable, given the widespread perception of cancer as a relentlessly progressive disease which, if left untreated, leads to death.

In fact, lesions called 'cancer' by pathologists can have very different growth rates. The concept of overdiagnosis is probably best understood by collapsing this spectrum of growth rates into four discrete categories, fast, slow, very slow and non-progressive, as depicted in Figure 38. Fast-growing cancers metastasize rapidly, produce symptoms and cause death. While they are potentially detectable by screening, they are easily missed and instead become evident in the interval between screening tests (so-called 'interval cancers'). Slowly growing cancers are destined to cause symptoms and death but can be detected by routine screening. It is on deaths from these cancers that screening is likely to have its greatest impact.

The two other growth rates represent cancers that never result in symptoms or death—cancers that Morrison (1992) and others have referred to as 'pseudodisease'. Some cancers progress so slowly that they are interrupted by death from unrelated causes before symptoms develop. The existence of this type of pseudodisease is therefore a function not only of the cancer's growth rate but also of the patient's competing risks for death. Although, in principle, screening will always lead to its detection (simply because some patients with screen-detected cancers will die of other causes), the problem is most relevant for cancer screening in the elderly, prostate cancer serving as the best example. Furthermore, some cellular abnormalities that are labelled 'cancer' never grow (or may even get smaller), and these non-progressive cancers will never cause symptoms, no matter how long

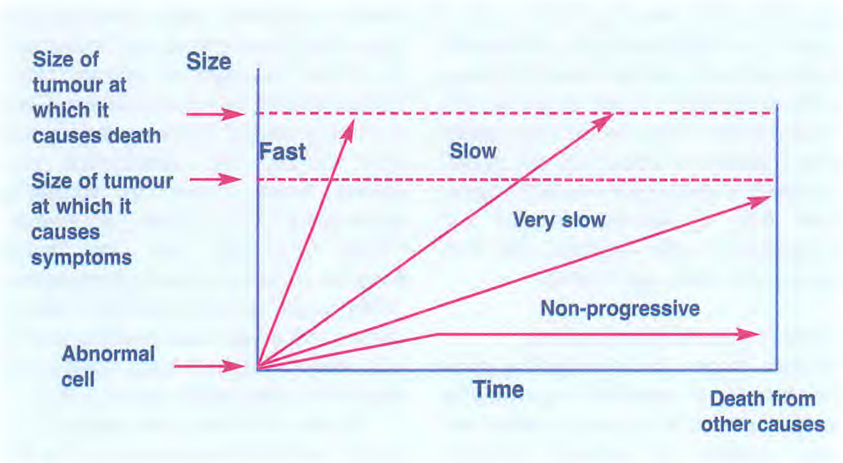
Size of tumour at which it causes death
Size
Fast
Slow
Size of tumour at which it causes symptoms
Very slow
Non-progressive
Abnormal cell
Time
Death from other causes

Figure 38 Growth rates of cancers

the patient lives. One can postulate a number of mechanisms for this type of pseudodisease: some cancers may outgrow their blood supply (and be starved), others may be recognized by the host's immune system (and be successfully contained), and some may never have been that aggressive. In the case of breast cancer screening, this second type of pseudodisease is probably the most relevant.

While the foregoing theoretical framework is a simplification of a wide spectrum of growth rates, it does serve as a basis for a straightforward definition of overdiagnosis. If pseudodisease is detected, then overdiagnosis has occurred. Practically, however, it is extraordinarily difficult to determine overdiagnosis, because virtually all detected cancers are treated, making it impossible to distinguish their natural history from the effect of treatment. Nonetheless, there are two sources of data from which some inferences about overdiagnosis can be made: randomized trials of mammography and population-based incidence rates.

Overdiagnosis in randomized trials of mammography

No screening test has been as thoroughly studied as screening mammography: over 500 000 women have been entered into eight randomized trials (see Chapter 4). As in each trial a group of women undergoing regular mammography (with or without clinical breast examination) is compared with those who are not, these studies provide some indication of the effect of mammography on the observed incidence of cancer, and the relative incidences in the screened and control groups can shed some light on the question of overdiagnosis.

The trials differ, however, in ways that potentially affect the rate of overdiagnosis. The screening intensity was greatest in the Canadian trials both because of the intervention (two-view mammography performed annually) and because of the high participation rate (nearly 100% at the first screening). If mammography resulted in overdiagnosis, it would be expected to be most obvious in these two trials. In trials with less intense interventions (single-view mammography every 2 or 3 years) or lower participation rates, overdiagnosis would be expected to be less evident.

Figure 39 shows the incidence of breast cancer in each of the trials at the end of the intervention period, 5–8 years after randomization. This figure highlights another characteristic that affects the relative incidence in the screened

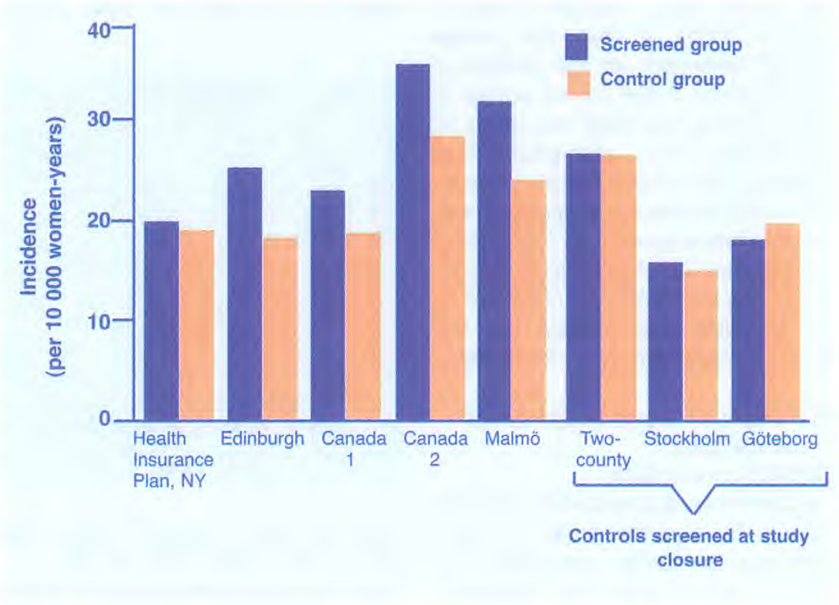

Figure 39 Breast cancer incidence among women invited to screening versus those who were not (controls) in eight randomized trials of screening mammography

and control group: whether or not the control group was screened at the close of the intervention period.

In the three Swedish trials in which the control group was screened at the end of the study, the relative incidence was essentially 1. This suggests either that there was no overdiagnosis or that it was confined to prevalent cases detected during the initial screening (their counterparts in the control group being detected during screening at closure). If there was overdiagnosis during subsequent screening, the control group would be expected never to catch up to a relative incidence of 1. The Two-county trial showed directly that the number of cancers detected after the prevalence round is no greater in the screened population than among controls.

In the five other trials, the relative incidence (screened versus control) ranged from 1.07 in the Health Insurance Plan trial to 1.38 in the study in Edinburgh. Because the Health Insurance Plan trial was performed in the 1960s, with mammographic equipment with much lower resolution than is available today, it does not provide relevant information on the problem of overdiagnosis. If the control groups in the remaining four trials are taken as representing the underlying 'true' incidence, then mammographic screening initially increases the observed incidence of breast cancer by 24–38%, suggesting potential overdiagnosis. Furthermore, long-term follow-up of women in the Canadian trials showed that this excess persists (Miller *et al.*, 2000, 2002).

Overdiagnosis in population-based incidence rates

Matched communities

Population-based data from two communities in The Netherlands point to overdiagnosis of similar magnitude and support the notion that the problem is largely confined to the initial screening (Peeters *et al.,* 1989b). In 1975, the City of Nijmegen started population-wide screening with mammography every 2 years. The neighbouring city of Arnhem, which had had a similar overall incidence in the preceding 5 years, served as the control. In the 4-year period immediately after initiation of screening, the overall incidence in Nijmegen was 30% higher than that in Arnhem. In the two subsequent 4-year periods, the incidence rates were again similar.

National screening programmes

Another means for investigating overdiagnosis is to examine breast cancer incidence rates in countries before and after initiation of national screening programmes. As other factors may influence incidence trends, inferences based on these data are less sure than those from randomized trials. Nevertheless, they offer the advantage of external validity; they offer, in fact, the best opportunity to see what happens in the real world. To make the inferences more secure, candidate countries should have initiated screening at a defined time, have programmes that are truly national in scope and have mature tumour registries. Two countries that meet these three criteria are Finland and the United Kingdom. In January 1987, Finland started two-view mammography screening every 2 years among women aged 50–59. The participation rate among those invited to screening approached 90% (Dean & Pamilo, 1999). One year later, the United Kingdom (Breast Screening Programme, 1999) began 3-yearly one-view (with a subsequent single view) mammographic screening of women aged 50–64 and reported a participation rate of > 70%.

In both countries, the incidence of breast cancer among women in the target age group rose after the introduction of screening (Figure 40). It should be emphasized that a temporary rise is not only expected but is necessary for screening to be successful, as the time of diagnosis is advanced for pre-existing cases (Morrison, 1992). While the rise may be temporary, it is nonetheless substantial: both countries experienced roughly a 50% increase in incidence in the target age group during 5 years after introduction of screening. Current data

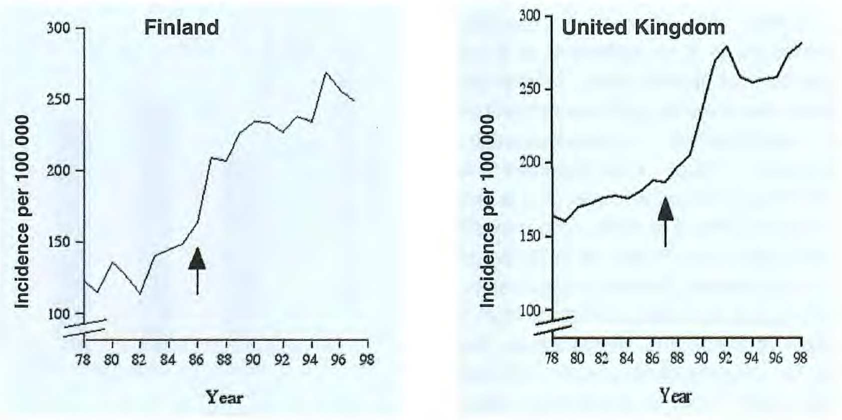

Figure 40 Trends in incidence of invasive breast cancer among women of screening age in Finland (50–59) and the United Kingdom (50–64)

Data from European Network of Cancer Registries (2001). The Finnish data are nationwide; data for the United Kingdom are from eight registries that have been collecting data since 1978 (East Anglia, Merseyside and Cheshire, North-western, South Thames, Trent, Yorkshire, Scotland and Wales). The arrow denotes the last data point before initiation of the national screening programme in each country.

from both countries suggest that the rise may be persisting.

Because incidence rises with age, one plausible explanation for the continuing rise in age-specific incidence is that the rates among women of screening age are ‘shifted up’ to the higher rates of older women. In other words, 60–64-year-old women assume the incidence rates of women aged 65–69 as their time of diagnosis is advanced. Were this to be the case, some fall in incidence would be expected in older, unscreened women. As shown in Figure 41, the breast cancer incidence in women aged 65–69 has in fact fallen slightly since 1991. Thus, to some extent, the increase in incidence simply reflects an advance in the time of diagnosis. However, the observed incidence rate among women aged 60–64 now exceeds that of women aged 65–69 and exceeds that which would be predicted in women aged 65–69 if the underlying 1.5% increase incidence had persisted. These data suggest that overdiagnosis is occurring in the United Kingdom. Early modelling indicated that overdiagnosis represents about 6% of detected cancers (Boer *et al.*, 1994).

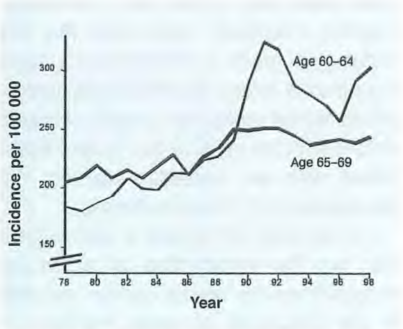

Figure 41 Age-specific breast cancer incidence trends in the United Kingdom for the oldest women invited to screening (60–64) and the next oldest unscreened group (65–69). Screening started between 1988 and 1995

Reservoir of potentially detectable breast cancer

In this section, we report on the ‘disease reservoir’ of breast cancer, which is the term given to the prevalence of disease observed at autopsy but undetected during life (McFarlane *et al.*, 1987). Although the evidence for this reservoir is derived from data on cancers detected after death, a proportion of those cancers could be detected during life, given enhanced imaging (e.g. computed tomography or magnetic resonance imaging) and/or more frequent biopsy. In other words, these data provide some sense of the upper limit of the problem of overdiagnosis of breast cancer.

A number of careful autopsy studies of the breast have been conducted (Table 55; Welch & Black, 1997). The series fall into two broad categories: hospital-based and forensic autopsies. The latter are consecutive cases presented to a coroner’s office (e.g. deaths in which homicide is suspected). Each series was restricted to women not known to have breast cancer during life. Although the level of scrutiny varied from study to study (e.g. in terms of how many tissue sections were made and whether post-mortem mammography was used), the same fundamental approach was used in each study, comprising systematic pathological examination of the breast.

Table 55. Studies of the prevalence of breast cancer in women not known to have had breast cancer during life, from autopsy series

Reference	Location	No. (autopsy type)	Mean no. of slides per breast	Invasive cancer (%)	DCIS (%)	Proportion of middle-aged women with any breast cancer (%)
Kramer & Rush (1973)	USA	70 (hospital)	40	1.4	4.3	ND
Wellings *et al.* (1975)	USA	67[a] (hospital)	ND	0	4.5	10 (age 50–70)
Nielsen *et al.* (1984)	Denmark	77 (hospital)	95	1.3	14.3	ND
Alpers & Wellings (1985)	USA	101 (hospital)	ND	0	8.9	13 (age 40–70)
Bhathal *et al.* (1985)	Australia	207 (forensic)	11	1.4	12.1	ND
Bartow *et al.* (1987)	USA	221 (forensic)	9	1.8	0	7 (age 45–54)
Nielsen *et al.* (1987)	Denmark	109 (forensic)	275	0.9	14.7	39 (age 40–49)

DCIS, ductal carcinoma *in situ*; ND, not described

[a] Reported as number of breasts, not number of women; prevalences are therefore percentages of breasts, not of women.

The median observed prevalence of invasive breast cancer among women not known to have breast cancer was 1.3% (range, 0–1.8%). The median prevalence of DCIS was 8.9%, but this varied widely: one series found none, while in three DCIS was found in over 10% of women undergoing autopsy. The observed prevalences were higher among women most likely to have been screened — middle-aged women — as much as one-third of whom showed some evidence of cancer. By comparison, the lifetime risk of dying from breast cancer was less than 4%. Consequently, many more breast cancers can be found than will ultimately matter to women.

Ductal carcinoma *in situ*

Aspects of the pathology and molecular biology of DCIS are described in Chapter 1. Clinical follow-up of women found to have DCIS has shown that it is a preinvasive neoplastic lesion and that the histological grade is related to prognosis (recurrence rates). Atypical ductal hyperplasia shows molecular genetic changes similar to those in DCIS and is also associated with an increased risk for the development of invasive cancer.

In the past, DCIS was a rare diagnosis, but the introduction of screening programmes for breast cancer resulted in the diagnosis of large numbers of cases of DCIS. High-grade DCIS more frequently shows abnormal mammographic features than low-grade DCIS, as the calcification present is more obvious, and the high-grade form is more specific for malignancy (see Chapter 1, Figure 42).

In screening programmes in Europe, the proportion of DCIS diagnosed ranged from 9 to 21% (Giordano *et al*., 1996; Fracheboud *et al*., 1998; Blanks *et al*., 2000a). In the USA, about one-third of all mammographically detected cancers are DCIS (Kerlikowske *et al*., 1993; Beam *et al*., 1996a; Poplack *et al*., 2000). With increased screening by mammography and increased sensitivity, perhaps combined with readier use of biopsy and diagnosis, the incidence (or, more correctly, the diagnosis) rates have increased dramatically. For example, the age-adjusted incidence of DCIS in the registries of the SEER programme in the USA have increased almost 10-fold over the past 20 years (from 2.7 to 25 per 100 000) (National Cancer Institute, 2001b).

Some researchers consider that detection of DCIS is one of the benefits derived from breast cancer screening. Indeed, aggressive screening for what was then called 'minimal breast cancer' was strongly advocated in the belief that only by detecting such lesions would mortality from breast cancer be reduced (Moskowitz *et al*., 1976). Minimal breast cancer as then defined consisted of two components: invasive breast cancers < 10 mm in size and DCIS.

Until recently, the usual treatment for DCIS was mastectomy. This probably explains the excess rate of mastectomy in the groups receiving mammography in the Canadian trials, for example (Miller, 1994), and the increased rate of mastectomy associated with increased detection of DCIS in the SEER programme (Ernster *et al*., 1996). With the advent of large numbers of mammographically detected DCIS, breast-conserving therapy has been used more widely. The two approaches have not been compared in a randomized controlled trial; however, when DCIS was treated by local excision, local recurrence was observed in 16% of cases within 4 years (Julien *et al*., 2000), and the percentage was significantly lower (9%) when radiotherapy was given. The risk for invasive recurrence was not related to the histological type of DCIS,

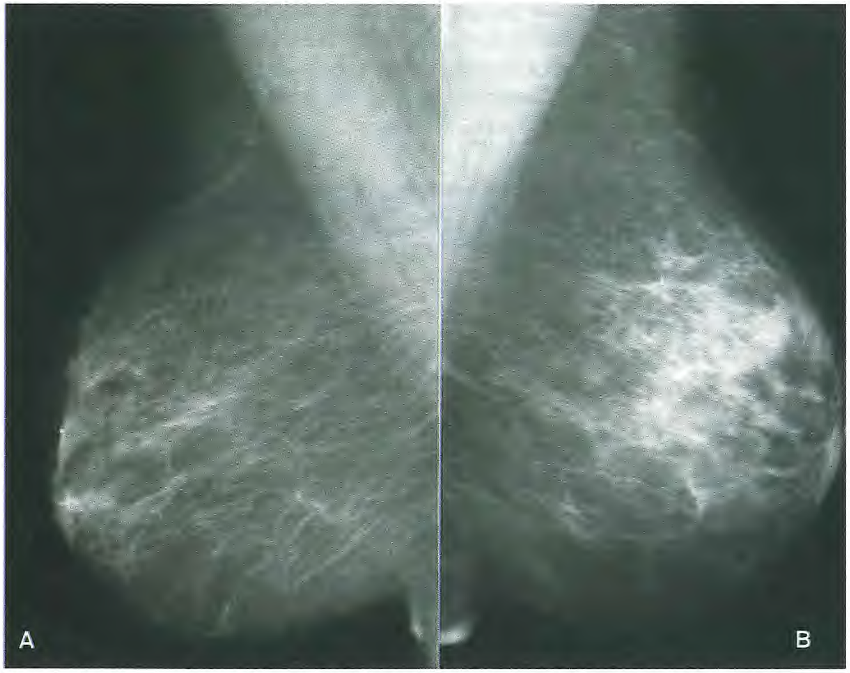

Figure 42 Mammograms of a patient who presented with a striking nonpalpable breast asymmetry. The structure of the tissue of the right breast (A) imitates glandular tissue. Magnetic resonance imaging (MRI) was performed, which showed intense enhancement with contrast medium of the patient's left breast (B). A histological diagnosis of high grade papillary ductal carcinoma in situ (DCIS) was made.

although distant metastasis was significantly more common in poorly differentiated DCIS (Bijker *et al.*, 2001b). About 20% of the DCIS lesions in these studies were palpable; the natural history of DCIS detected solely by mammography is less clear. Holland *et al.* (1990) found no association between the mode of detection of DCIS and the size of the resected lesion. Frequently, DCIS extends over more than one quadrant of the breast, making breast-conserving surgery impossible.

The important issue is what proportion of cases of DCIS detected at screening would have progressed to invasive cancer if they had not been detected, and how many represent 'overdiagnosis'. If a substantial proportion of cases of DCIS were destined to progress to invasive breast cancer, some decrement in the rate of invasive breast cancer would be expected as DCIS was increasingly diagnosed (a 'new case of cancer' can occur only once). In other words, as DCIS becomes an increasingly common diagnosis, women destined to get invasive cancer are counted as new cases of DCIS, not as new cases of invasive breast cancer. One diagnosis is substituted for the other.

Information on the extent to which the incidence of invasive breast cancer might be reduced by detection of DCIS is available from the long-term follow-up in the Canadian National Breast Screening Study of women aged 50–59 on entry (Miller *et al.*, 2000). Of the 267 invasive breast cancers detected at annual mammography, 48 were < 10 mm in size, whereas only 6 of 148 found by clinical breast examination were this size. In addition, 71 in situ breast cancers were detected in the women receiving annual mammography and 16 in those examined physically. However, there was no evidence that the detection of in situ cancers resulted in a reduction in breast cancer incidence: the cumulative numbers of invasive breast cancers (including those ascertained after the end of the 4–5-year screening period) were 622 in women with annual mammography and 610 in those given clinical breast examination. The data for the 50 000 women aged 40–49 on entry to the Canadian trial are similar (Miller *et al.,* 2002). Once again, more in situ cancers were diagnosed in women given mammographic screening (71 cancers) than in those receiving usual care (29 cancers). However, no indication was found of a reduction in breast cancer incidence over the 11-year follow-up, the cumulative numbers of invasive cancers as determined by linkage to the national cancer registry being 592 and 552, respectively.

Similar follow-up data have not been published from the other breast cancer screening trials. However, many have reported the proportion of DCIS among the cancers detected, ranging from 8.4% in the Two-county trial to 16% in the Malmö trial (Fletcher *et al.,* 1993).

Studies of populations in which breast cancer screening has been implemented provide no evidence that the rising rates of incidence (diagnosis) of DCIS have been accompanied by a decrease in the incidence of invasive cancer. Data from the SEER programme in the USA (Figure 43) indicate that detection of DCIS simply adds to the total number of cancers detected (although, of course, the increase in incidence of invasive cancer might have been even greater in the absence of DCIS detection and treatment).

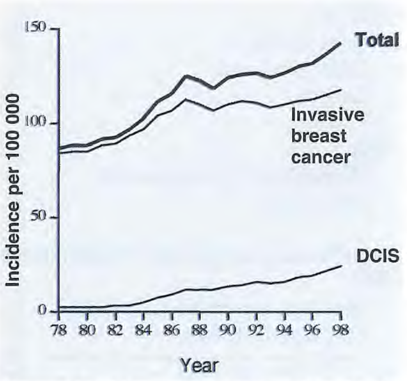

Figure 43 Incidences of ductal carcinoma *in situ*, (DCIS), invasive breast cancer and the two combined, USA, 1978–98
From National Cancer Institute (2001b)

DCIS does not always progress to invasive breast cancer, as shown by two lines of evidence. The first derives from studies of cases of DCIS followed only by biopsy (without treatment). Of 28 cases of non-palpable, low-grade (non-comedo) DCIS, seven (25%) developed into invasive cancer within 10 years and a further two within the next 20 years (Page *et al.*, 1995). In Bologna, Italy, only 3/28 cases (11%) of DCIS developed into invasive cancer during an average follow-up period of 17 years (Eusebi *et al.,* 1989). The second type of evidence derives from studies of the outcome (survival) of registered DCIS cases. The SEER data for 7072 cases of DCIS registered in 1978–89 showed a low risk for death from breast cancer during the 10-year follow-up. In cases diagnosed in 1978–83 (mainly before mammographic screening), the risk for death from breast cancer was 3.1 times (at 5 years) and 3.4 times (at 10 years) that of the general population. In cases diagnosed in 1984–89 (mainly detected by mammography), the relative risk for death was much lower: 1.6 (95% CI, 1.1–2.1) at 5 years and 1.9 (95% CI, 1.5–2.3) at 10 years. Some of the excess risk may be due to 'missed' invasive disease among the DCIS cases. The women with DCIS had an overall mortality rate that was 20–30% lower than that of the general population, as they represented a group with higher socioeconomic status, who undergo frequent mammography (Ernster *et al.*, 2000).

Part of the difficulty in determining the role of DCIS in screening is the fact that mammography has revealed a new spectrum of disease that would have been largely undiagnosed in the absence of screening, although the presence of a 'reservoir' of DCIS was evident from studies of autopsy series (see above).

This has contributed to confusion as to what these lesions truly represent and even whether it is appropriate to use the term 'carcinoma' when no precise guidance can be provided on eventual prognosis (Foucar, 1996). Miller and Borges (2001) suggested that true precursors of invasive cancer, with atypical epithelial hyperplasia and incipient invasion, are not detectable with current screening methods. Some high-grade cancers may have a transitory in situ phase, with rapid progression to invasion, thus not allowing time for their detection as DCIS (Barnes *et al.*, 1992).

Early mortality from breast cancer

A possible, nonsignificant excess of early mortality from breast cancer was noted among groups invited to screening, especially women under the age of 50, in some screening trials (Tabár *et al.,* 1985; Andersson *et al.*, 1988; Miller *et al.*, 1992a, 2002). In a meta-analysis of all screening trials with data on women under the age of 50, Cox (1997) noted that the rate ratio for mortality at 3 years was 2.4 (95% CI, 1.1–5.4), although all the rate ratios computed earlier or up to 11 years after entry were approximately 1.0. Miettinen *et al.* (2002), in an analysis of the data on women aged 55–84 on entry to the Malmö trial, noted a 3-year average rate ratio of 1.5 3–4 years after entry, although the lower 95% CI was approximately 0.6 (data derived from a figure). Retsky *et al.* (2001a,b) suggested that the surgical removal of a primary breast tumour from premenopausal women with involved lymph nodes triggers the growth of temporarily dormant disease in approximately 20% of cases. This suggestion is in line with the results of tests in experimental animals (Fisher & Saffer, 1989).

Risk for breast cancer induced by radiation

Several epidemiological studies have addressed the risk for breast cancer induced by radiation and provided quantitative estimates of the level of risk after different doses of radiation. Age-related risk estimates are needed for calculating the risk of a patient undergoing mammography, as the risk for radiation-induced breast cancer decreases with age at exposure. Additionally, information about the time between exposure and diagnosis is essential for risk evaluation, as there is a latency of 5–10 years between irradiation and the appearance of any excess cancer risk, with further increases over the next 5–10 years. The risk then probably persists for the remainder of the lifespan. The latency may be longer with lower doses. Although the data are not fully consistent, dose fractionation does not appear to reduce the risk. In other words, the effect of low doses seems to be additive, and low-dose fractions appear to be as effective in inducing breast cancer as a single large dose.

Age-specific risk estimates from large epidemiological studies

Sixteen epidemiological studies have provided relative risk estimates for breast cancer associated with exposure to radiation: 12 for incidence and four for mortality (for overviews, see Boice, 2001; Little, 2001). Only eight of the studies of incidence and three of those of mortality provide estimates for women aged ≥ 20 at the time of exposure. Age-specific excess relative risks (ERRs) were summarized by UNSCEAR (1994, Annex A) and are presented in Table 56.

Table 56. Excess relative risks for death from breast cancer associated with exposure to radiation by age at time of exposure

Study	Age at exposure (years)			
	20–29	30–39	40–49	> 50
Japan, 1950–85				
Incidence	1.27	1.23	0.54	0.31
Mortality	0.96	1.09		
Sweden, treatment for benign breast disease	1.9	0.4	0.1	0.1
Canada, treatment for tuberculosis, 1950–80				
Nova Scotia	1.6	0.8	0	0
Other provinces	0.4	0.2	0	0
Canada	0.22	0.04	–0.01	–0.03
USA				
Massachusetts, treatment for tuberculosis	0.5	0.2[a]	0.2[a]	0.2[a]
California, treatment for Hodgkin disease	0.4[b]	0.4[b]	0.1[b]	0.1[b]
New York, treatment for mastitis	0.4	0.6		
Contralateral breast	–	0.2[c]	0.2[c]	0.0

From UNSCEAR (1994; Tables 11 and 20)

[a] One risk coefficient reported for women aged ≥ 30 at exposure

[b] One risk coefficient reported for women aged 20–39 and one for women aged ≥ 40 at exposure

The two largest cohort studies are the Life Span Study cohort of survivors of the atomic bombs in Japan and a cohort study of women who were examined frequently by X-ray fluoroscopy during therapy for tuberculosis in Massachusetts, USA. Age-specific risks were also reported in a Canadian study of women with tuberculosis and a Swedish study of benign breast disease. Other studies carried out on women irradiated for diagnostic, therapeutic or occupational reasons are summarized in *IARC Monographs* Volume 75 (IARC, 2000).

Risk models for dose–response analysis

The linearity of the dose–response curve for the ERR (or RR = 1 + ERR) is widely accepted for the relationship between exposure to radiation and risk for breast cancer. Two models, an 'age-at-exposure' model and an 'age-attained' model have been described. The 'age-at-exposure' model is often presented as a function of the form:

$$\text{ERR}\ (D) = D \times \alpha \times \exp(-\ \beta\ (e - 30)),$$

where *e* is the age at time of exposure, *D*, the dose of radiation and α can be interpreted as the excess relative risk (ERR per Sv) of a women exposed at the age of 30. The model for the 'age-attained' model is often given in the form:

$$\text{ERR}\ (D) = D \times \alpha \times \exp(-\ \beta\ (a - 50)),$$

where *a* is the age at diagnosis and α can be interpreted as the excess relative risk (per Sv) of a woman aged 50.

In models fitted with data from the Life Span Study, the ERR increased by 3.7% per year with the 'age-at-exposure' model, whereas the ERR increased by 4.6% per year with the 'age-attained' model. No significantly increased risk was seen for women aged ≥ 50 at the time of exposure (Thompson *et al.,* 1994; Tokunaga *et al*., 1994).

The Committee on the Biological Effects of Ionizing Radiation (BEIR; 1990), the International Commission for Radiation Protection (ICRP, 1982, 1987, 1991) and the National Council on Radiation Protection and Measurements (USA; 1986) have reported risk coefficients for various organs, but the values are difficult to compare. The BEIR Committee reported age-specific coefficients for breast cancer separately. The lifetime risks for death due to radiation-induced breast cancer were given as 0.43% per Sv for women aged 30–40 at the time of exposure, 0.2% per Sv for women aged 40–50 and 0.06% for women aged 50–60. The risk coefficient for women > 60 at the time of exposure was 0. Most predictions of radiation-induced breast cancer are based on the age-specific risks reported by the BEIR Committee. Jung (2001) presented the risk coefficients reported by the three committees in a comparable way and showed that they differed substantially, particularly for exposure after the age of 50.

Dose of radiation from mammography

Considerable effort has been expended in estimating and measuring the dose of radiation to the breast during mammography (Hammerstein *et al.*, 1979; Stanton *et al.,* 1984; Wu *et al*., 1994; Young *et al.*, 1996; Young & Burch, 2000). As it is now generally assumed that glandular tissue is the most vulnerable of the tissues making up the breast, some authors consider that the average dose to the glandular tissue is the most appropriate dosimetric quantity for predicting the risk for cancer. This quantity is also recommended by the ICRP (1987) and others. The average glandular tissue dose is calculated by multiplying the measured 'entrance surface air kerma free in air' by *g*, a conversion factor, which depends mainly on the radiation quality and the thickness and tissue composition of the breast.

The exposure of two groups of women who underwent mammography at a gynaecological clinic during two different periods was compared. The first group comprised 1678 women who were examined between February 1992 and July 1992 with a tungsten and wolfram anode tube, and the second comprised 945 women who were treated 1 year later (July 1993–November 1993) with a dual-track molybdenum–wolfram anode mammographic unit. The mean average glandular tissue doses were 1.6 and 2.1 mGy, with an average of 3.4 and 3.6 exposures per women, respectively (Klein *et al.,* 1997).

The doses to other organs were considered by the ICRP (1982) and were assumed to be negligible. Since then, few measurements of doses to other organs have been undertaken. In a report on thyroid doses due to mammo-graphy in 91 women, the average dose to the skin overlying the thyroid was 0.39 mGy per mammographic examination, the values ranging from background levels to 1.2 mGy. The authors estimated that this value corresponds to an average dose to the thyroid during mammography of 0.04 mGy, with an average dose to the breast of 4 mGy (Whelan *et al.,* 1999)

In a study of the dose absorbed by seven organs other than the breast during mammography, the absorbed dose was measured with an anthropomorphic phantom containing thermoluminescent dosimeters. Doses to the red bone marrow of the sternum and the thyroid, lung, liver, colon, oesophagus and stomach were considered. The mean dose to the red bone marrow was 0.40–1.3 μGy/mAs, and that to the thyroid was 0.05–0.17 μGy/mAs. The doses to the other five organs were considered negligible. When the effective dose of radiation was calculated, the dose to the breast contributed > 98% (Hatziioannou *et al.*, 2000).

Estimated numbers of cases of or deaths from breast cancer due to mammography

The risks and benefits of specific screening policies that include mammog-

Table 57. Estimates of numbers of cases of breast cancer induced by screening mammography, with assumptions about screening policy, radiation dose and age at screening

Reference, country	Dose, view	Total dose per examination (both sides)	Total dose per length of observation	Screening policy, age at screening	Breast cancer incidence per 10^6 women/mGy	Breast cancer deaths per 10^6 women/mGy	Lifetime risk per 10^6 women	Comments
Howe *et al.* (1981), Canada	0.7 mGy; max, 2.5 mGy		10 mGy 25 mGy 5 examinations	Age 40–59, annually for 5 years	0.66 0.87			Two values are for linear additive model and linear multiplicative model plus term for cell killing Latency, 10 years Mortality reduction, 40%
Feig (1984), USA		1–8 mGy		Age at exposure:	0.35	0.175		Latency, 10 years National Cancer Institute model
				40			10	
				45			9	
				50			7	
				55			6	
				60			4	
				65			3	
				70			2	
Zuur & Broerse (1985), Netherlands		1 mGy	30 mGy/30 examinations	Age 35–75, annually	20	8	600 cases and 240 deaths 1071 cases and 428 deaths	Absolute risk model Relative risk model Latency, 10 years
Gohagan *et al.* (1986), USA	0.6 mGy 4 mGy	2–3 views, 1.2–1.8 mGy 2–3 views, 8–12 mGy		Age 35 at baseline examination; age 40–80 annually	2–3		150 cases with 1.2 mGy/examination; 1000 cases with 12 mGy/examination	Low-dose film–screen system vs 4-mGy system; breast compressed to 6 cm
Law (1987), United Kingdom	1 mGy 10 mGy	2 mGy 20 mGy		Age at exposure: 35 65 Single examination at each age			18.6 3.6	Breast compressed to 5 or 8 cm at 2 doses Latency, 10 years
Hasert (1988), Germany		10 mGy 1 mGy		Age ≥ 35, annually	0.35 0.35	0.18 0.17		Not screen–film Screen–film Latency, 10 years

Table 57 (contd)

Reference, country	Dose, view	Total dose per examination (both sides)	Total dose per length of observation	Screening policy, age at screening	Breast cancer incidence per 10^6 women/mGy	Breast cancer deaths per 10^6 women/mGy	Lifetime risk per 10^6 women	Comments
Mettler *et al.* (1996), USA	1.38 mGy	2.8 mGy	112 mGy/40 examinations 98 mGy/35 examinations 70 mGy/25 examinations	Age 35–75 Age 40–75 Age 50–75 Annually			15 10 5	Mortality reduction: 35–39, 5% 40–49, 15% 50–75, 25% Mortality, 40% Average lifespan, 75 years Latency, 10 years
Andersson & Janzon (1997), Sweden	2 mGy	4 mGy	36 mGy/9 examinations 20 mGy/5 examinations	Age, < 50 Every 2 years		5		Mortality reduction, 36% (adjusted for fatal radiation-induced breast cancer)
Feig & Hendrick (1997), USA	2.5 mGy 5.5–6.5 mGy	4 mGy		Age 40 Annual screening for 10 years Screening every 2 years for 10 years		0.05 0.05	8 deaths 4 deaths	Breast compressed to: 4.2 cm 5–5.7 cm
Beemsterboer *et al.* (1998a), Netherlands	2 mGy	1 view, 2 mGy 2 views, 4 mGy	2 views at first screening, 1 view subsequently: 22 mGy 32 mGy 42 mGy 62 mGy	Age 50–69, every 2 years Age 40–69, every 2 years Age 40–49 annually and 50–69 every 2 years Age 40–69 annually				Dose measured on phantom Attendance rates: 40–49, 75% 50–69, 70%
Mattson *et al.* (2000), Sweden	1.5 mGy	Average, 2.25 mGy 1.5 views/ examination	13 mGy/7 examinations 33 mGy/17 examinations	Age 40–49, every 18 months Age 50–69, every 2 years			Incidence: 530 with assumption of higher risk 120 with assumption of lower risk Deaths: 240 (higher risk) 50 (lower risk)	Annual mortality reduction beginning 7 years after examination: 40–49, 25% 50–69, 30% Latency, 10 years Attendance rate, 80% Recall rate, 5%, with 3 views at recall examination

Table 55 (contd)

Reference, country	Dose, view	Total dose per examination (both sides)	Total dose per length of observation	Screening policy, age at screening	Breast cancer incidence per 10^6 women/mGy	Breast cancer deaths per 10^6 women/mGy	Lifetime risk per 10^6 women	Comments
Young & Burch (2000), United Kingdom	Average mean glandular tissue dose: 2.03 mGy (oblique view) 1.65 mGy (cranial-caudal view)	3.68 mGy	2 views at first visit: 18.4 mGy; 1 view (oblique) at subsequent 4 visits: 11.8 mGy					Breast compressed to 4–4.5 cm
Jung (2001), Germany	Mean parenchymal dose/-view, 2 mGy	4 mGy	24 mgGy/6 examinations	Age at first exposure: 37.5 42.5 52.5 57.5 62.5 Every 2 years	13 7.1 4.8 1.8 0.95	5.5 3.1 2.1 0.79 0.42	Deaths: 133 74 51 19 10	Breast compressed to 5.0–5.5 cm Relative biological effectiveness, 2 Equivalence dose, 8 mSv Latency, 12 years
Säbel *et al.* (2001), Germany			2 mGy	Age 40–49 Age 50–59	4.5 1.5	2.0 0.65		Relative biological effectiveness, 1 Morbidity:mortality, 2.3

raphy have been estimated in a number of publications, sometimes with estimates of the numbers of cases of or deaths from radiation-induced breast cancer. However, different values for radiation dose, different risk models and different assumptions about age range and screening interval in the mammography programmes were used in the various papers. Some provided only risk–benefit ratios under an assumption for the effect of mammography screening and did not provide the numbers of radiation-induced breast cancer cases or deaths. Others calculated various indices for possible harm due to radiation and estimated, e.g., the lifetime risk of 1 million women. The model-based calculations are difficult to compare, as the results are presented differently. The assumptions used and the estimates made are summarized in Table 57.

Howe *et al.* (1981) assumed that women aged 40–59 were screened five times with a dose of 0.7–2.5mGy per view and used various models to calculate the number of induced cancers. A mortality reduction factor of 40% was taken to calculate the number of deaths from breast cancer; however, the estimated number of deaths was not given, as only the combination of induced and 'saved' deaths was calculated. In the model, 553 deaths from breast cancer would have occurred among unscreened women 20 years after entry into the study, while 487 (additive model) or 490 (multiplicative model) women in the screened group would have had breast cancer. The corresponding numbers after 30 years were 892 deaths in the unscreened group and 825 (additive model) or 831 (multiplicative model) in the screened group. It was concluded that the number of radiation-induced cancers is negligible in comparison with the spontaneous incidence.

Zuur and Broerse (1985) calculated the risk for breast cancer of women aged 35 who were screened with 1 mGy per examination every year until the age of 75, that is, 40 times. A latency of 10 years was assumed. Models for absolute and relative risk were used. With the absolute risk model, the estimated lifetime risk for 1 million screened women was 600 induced cases of breast cancer (incidence) and 240 deaths. The relative risk model resulted in somewhat larger numbers: 1071 additional breast cancer cases and 428 deaths from breast cancer.

Gohagan *et al.* (1986) estimated the number of deaths from breast cancer induced by a screening policy that included one baseline examination at the age of 35 and an annual examination between the ages of 40 and 80. Two mammographic techniques were assumed: a low-dose film–screen system emitting 0.6 mGy dose per view and a system emitting 4–mGy per view. Two to three views were assumed at each examination, resulting in typical absorbed doses of 0.12 and 0.18 mGy, or 8–12 mGy. In a linear dose–response model, the lifetime radiogenic risk in a population of 1 million women screened was 150 (low-dose film) or 1000 breast cancer cases (4-mGy system), compared with 93 000 'spontaneous' cases.

Law (1987) investigated the effect of a programme in which women were screened between the ages of 35 and 65 in three risk models. Screening at the age of 35 resulted in 18.6 additional cases (per examination), while screening at the age of 65 gave 3.6 induced cases per million screened women. The author concluded that screening from the age of 35 with current techniques was not recommendable.

Hasert (1988) compared the numbers of radiation-induced breast cancer cases that would be induced with a dose of 10 mGy per examination in conventional techniques and 1 mGy with a new screen–film combination. For women over 35, he estimated that there would be 3.5 additional cases of breast cancer per million women exposed to 10 mGy and 0.35 additional cases with the lower dose.

Mettler *et al.* (1996) assumed annual mammography beginning at 35, 40 and 50 years and continuing until the age of 75. The dose at each examination was assumed to be 2.8 mGy, resulting in a total dose of 112 mGy if screening began at the age of 35, 98 mGy with screening from the age of 40 and 70 mGy with screening from the age of 50. A linear model from the Life Span Study and a latent period of 10 years were used. If mammography was started at the age of 35, 15 fatal induced breast cancer cases per million women were predicted, 10 if screening started at the age of 40 and five if screening started at the age of 50. The authors also calculated risk–benefit ratios for breast cancer screening, arguing that the benefit to a woman beginning annual screening at the age of 35 and continuing until 75 would be 25 times greater than the potential risk. If screening began at the age of 50, the risk–benefit ratio would be about 100.

Andersson and Janzon (1997) assumed a dose of 4 mGy for each examination (both sides), resulting in a total dose of 36 mGy for women undergoing nine examinations and 20 mGy for women undergoing only five examinations. They assumed that screening every 2 years started at the age of < 50 but with incomplete participation rates. The calculations are based on a linear dose–response model with age-dependent risk coefficients. Ten radiation-induced breast cancer deaths were estimated per million women screened.

Feig and Hendrick (1997) assumed screening annually or every 2 years, beginning at the age of 40 or 50, and estimated the numbers of radiation-induced breast cancer with a dose of 4 mGy at a two-view examination. Three models were used to determine the dose-–response relationship, with assumptions of a 10-year latency and age-specific factors from the BEIR Committee (BEIR, 1990). Annual screening for 10 years from the age of 40 was estimated to result in eight deaths from breast cancer per million women (lifetime risk), while screening every 2 years resulted in four induced breast cancer deaths. In an earlier paper, Feig (1984) compared linear, linear–quadratic and quadratic dose-–response models, with an assumed latency of 10 years. In their worst-case scenario (assuming a dose of 100 mGy), 20 excess deaths from breast cancer would be induced during a lifetime. Fewer induced breast cancer cases were estimated with the other models.

Beemsterboer *et al.* (1998a) estimated the number of breast cancer deaths among 1 million screened women induced by various mammography programmes. The latency was taken to be 10 years, and models based on age at exposure and attained age were used with coefficients calculated by the BEIR Committee. It was further assumed that the ratio of incidence to mortality rates is 2.6. The total exposure of women screened every 2 years between the ages of 50 and 69 was estimated to be 22 mGy. With these assumptions, 5.1 deaths from breast cancer were estimated to be induced. With screening every 2 years between the ages of 40 and 69, for a total of 30 examinations, 7.3 deaths were estimated to be induced per 1 million women. For screening every 2 years between the ages of 50 and 69, the baseline scenario, the ratio of induced:prevented breast cancer cases was estimated to be 1:242, whereas the ratio was 1:97 when screening was performed for women aged 40–49 at a 2-year interval and 1:66 at a 1-year interval.

Mattsson *et al.* (2000) compared the risk–benefit relationship for a reduction in breast cancer mortality in various models and with various assumptions. Two polices were compared: screening of women aged 40–49 at an 18-month interval and screening of women aged 50–69 every 2 years, which would result in lifetime doses of 13 and 33 mGy, respectively. Risk models from various epidemiological studies were used. In a hypothetical cohort of 100 000 women aged 40 who were followed-up until the age of 100, the number of induced deaths ranged from 5 to 24 and the number of years lost from 71 to 325.

Jung (2001) investigated the risk of mammography in two models of screening: screening every 2 years and screening starting at different ages but continuing for 10 years for a total of six examinations. He assumed a mean parenchymal dose per view of 2 mGy, resulting in 4 mGy per examination and thus 24 mGy from the six examinations of the screening programme. He also assumed a relative biological effectiveness of 2, and consequently a dose of 8 mSv per examination, and a linear dose–response model based on BEIR Committee coefficients (BEIR, 1990). The risk–benefit ratio for women first screened in their 40s was about 6, and that for women first screened in their 50s was about 25. The risk for developing breast cancer increased from 9 for unscreened women to 9.036, and the risk for dying from breast cancer increased from 3.96 to 3.961. He concluded that the risk for death from breast cancer is negligible if screening starts at 50 but should be taken into consideration in screening women aged 40–50.

Säbel *et al.* (2001) calculated the risk associated with a single examination at 2 mGy for women aged 40–49 or 50–59. In the younger women, 4.5 cases of breast cancer would be induced per 1 million women, while for the group aged 50–59 only 1.5 additional cases would be induced. The incidence:mortality ratio was taken as 2.3, resulting in 1.96 and 0.65 breast cancer deaths, respectively.

Although the authors of these studies use different assumption for the screening programmes, such as different age groups, screening intervals, doses of radiation at each mammography and models to estimate the numbers of radiation-induced breast cancer cases, the results are consistent in showing that few breast cancer cases are induced by radiation during mammography. If screening was begun at the age of 50, the number of deaths from breast cancer during the remaining lifespan was estimated to be 10–50 per million regularly screened women (10–20 screens, 2–5 mGy per screen), while if regular screening was begun at the age of 40, the number of radiation-induced deaths from breast cancer would be 100–200. These numbers can be compared with the tens of thousands of breast cancer deaths in un-screened populations (cumulative mortality). The low additional risk is due to the fact that exposure to radiation after the menopause is associated with a low risk, as observed in many epidemiological studies.