FROM EVIDENCE TO DECISION SUPPORT IN CANCER SCREENING

applications of Miscan models

VAN EMPIRISCHE KENNIS NAAR BELEIDSONDERSTEUNING BIJ KANKERSCREENING
toepassingen van Miscan modellen

Proefschrift
ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof.dr. P.W.C. Akkermans M.A.
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
woensdag 28 juni 2000 om 13:45 uur

door
Robert Boer
geboren te Vlaardingen

Promotiecommissie

Promotor:	Prof. dr. P. J. van der Maas
Overige leden:	Prof. dr. J. D. F. Habbema
	Prof. dr. M.G.M. Hunink
	Prof. dr. E.K.A. van Doorslaer
Copromotor:	Dr. ir. G. J. van Oortmarssen

ISBN 90-72245-94-6

This thesis is partly realised due to the financial support of the Department of Public Health

CONTENTS

1. INTRODUCTION

Making decisions to introduce or change cancer screening that are based on observed evidence is not straightforward. Even if there is evidence available from randomised controlled trials that clearly proves a reduction in mortality from the cancer in question, there are still several issues to be resolved before an evidence based decision can be made on the introduction or change of a screening programme. Screening inevitably comes with unfavourable effects such as extra cancer incidence with the burden of treatment of those cancers and its after-effects, and the burden from an increase of diagnostic procedures and of the screening procedure itself. This raises the question whether favourable effect of screening (mortality reduction, prevention of advanced disease) are sufficiently large in comparison to the unfavourable effects. Other questions are: How costly is screening? How are the observed effects to be extrapolated to a situation that is different with respect to demography, epidemiology and health care? There are also several optimisation questions. What is the optimal screening test? What is the optimal way of organising the screening that is to be offered to the public? What is the optimal screening schedule with respect to ages and intervals between screenings? Each of these questions concerns a trade-off of favourable and unfavourable effects of screening that may turn out differently for the screening programme that is considered than for the situation in which empirical evidence was observed. Thus there are several steps to be made in between the empirical investigation that produces observed evidence concerning cancer screening and the decision to introduce or change cancer screening. This thesis concerns neither empirical investigation nor decision-making, but the intermediate steps between the two. The steps that are being addressed are the gathering of evidence, the uncertainty associated with present evidence, the balancing of favourable and unfavourable effects and the influence of particular circumstances in which cancer screening takes place. Finally there is the possible need for side steps to go into compelling questions that arise along the way. The examples chosen to illustrate these steps are part of the work of the research group on screening evaluation of the department of Public Health of Erasmus University Rotterdam.(Beemsterboer 1999; de Koning 1993; Koopmanschap 1994; van Ballegooijen 1998b; van Oortmarssen 1995; Wildhagen 1999)

Most examples rely to a large extent on the application of Miscan simulation models. (Habbema et al. 1985; Loeve et al. 1999). The Miscan program simulates a series of individual life histories considering date of birth and of death, development of the disease in question before diagnosis in a situation without screening, diagnosis and survival, and the influence of a screening programme on the date and stage of detection and the date of death. A life history is simulated through individual realisations of specified probability distributions concerning demography, the course through a number of possible disease states, sojourn times in these disease states, behaviour with respect to the screening programme, and test characteristics and consequences of screening. The life histories are aggregated to a population in which incidence, life years with disease, life year lost from disease, numbers of screening, stage distribution at diagnosis etc. are counted. A Miscan model can reproduce the circumstances under which empirical evidence has been gathered, can compare observed and modelled results so that one can check to what extent the model assumptions on natural history can explain what is observed. Likewise, a Miscan model can reproduce the circumstance of a future screening programme for which the results can be predicted and thus support decision making.

gathering evidence

Observed evidence on medical interventions is usually generated in circumstances that are rather different from future daily applications of that intervention. That also means that, in order to predict effects under those different circumstances, there is need for evidence from different types of studies.

Randomised trials in cancer screening aim at testing whether screening results in a reduction in mortality from the cancer. In doing so, these trials also produce an estimate of the size of the mortality reduction. But since the size of a trial is usually chosen to be just sufficient to produce a mortality reduction that is just significantly different from 0, the confidence interval for an estimate of the size of the mortality reduction is wide. A more precise estimate can be derived by combining the results of more than one trial. For this the concept of meta-analysis has been developed. However, a meta-analysis gives a more precise estimate at the expense of the intervention being defined less precisely.(Blettner et al. 1999; Davey Smith and Egger 1998; Davey Smith et al. 1997) In an attempt to resolve this problem, a model can be used that is structured to explain the effects of screening. The direct estimate of mortality reduction is substituted by an estimate of crucial parameters that explain the amount of mortality reduction in a screened population. The model can apply the same assumptions on natural history and effects of early detection by screening to the settings of the different trials and thus regain precision of the definition of the intervention. An example of such parameter

estimation is presented and discussed in chapter 2 of this thesis.
Randomised trials are sometimes practically not feasible and sometimes they are just not the most appropriate method for acquiring knowledge. But observational studies are prone to bias. In real life it is usually not possible to investigate bias in observational studies. Or rather, as far as bias can be determined in observational studies, that bias can also be eliminated. Therefore the remaining bias cannot be investigated in real studies. We applied micro-simulation models that were originally developed for evaluation of screening programmes to investigate sources of bias in observational studies.
Case-control studies are used to estimate mortality reduction due to cancer screening. The general design of case-control studies is usually adjusted for estimating efficacy of cancer screening. Within each set of a case who dies from the cancer and its matched controls, exposure to screening among controls that occurs later than the diagnosis with the disease in the case is disregarded.(Cronin et al. 1998; Sasco et al. 1986) This is to compensate for bias due to the fact that after diagnosis one is not screened any more. Chapter 3 shows that this is overcompensation and still leads to bias. It also shows that there are several other serious biases possible if the particulars of the timing of screening in the population under study are not carefully considered. These biases occur next to bias due to the association of risk for the disease and the individual tendency to participate in screening.
Besides case-control studies, that can be seen as weak alternatives for randomised controlled trials, there are several other types of estimates that can be of value for evaluating cancer screening. An important mediator of cancer screening effects is net survival from the disease. Net survival shows the mortality effect among individuals with the disease that is attributable to that disease.(Estève et al. 1994) Estimates of net survival can be biased in many ways. Chapter 4 investigates an alternative for the standard methods of net survival estimation. This retrospective survival selects the population that is used for the estimate from the people who have died in a certain, relatively small, period. In contrast, usually survival estimates select the population from newly diagnosed cases of the disease. Retrospective survival is shown to possibly result in large bias. In chapter 5 it is shown for the colorectal cancer and prostate cancer cases in the SEER program that different standard methods of survival estimation, do not result in very different outcomes.

evidence and uncertainty
As mentioned earlier, empirical evidence on essential aspects of cancer screening, such as its quantitative influence on mortality from the disease, tends to be not very precise. Besides the question of how to establish more precision, there is also reason for explicit concern about the propagation of uncertainty on different aspects of cancer screening into conclusions for decision support. The most vigorous method to describe this propagation is an uncertainty analysis.(Cox and Baybutt 1981; Morgan and Henrion 1990) For this type of analysis it is assumed that the uncertainty about the true model parameters is represented by a probability distribution that leads to a probability distribution in model outcomes, which in turn is assumed to represent uncertainty in the model outcomes. Such analysis is not part of this thesis.(Chessa et al. submitted) But in several chapters sensitivity analyses are applied in which model scenarios with different values for uncertain model assumptions are evaluated in order to study the possible effects of these values being overestimated or underestimated. Chapter 6 shows a sensitivity analysis concerning two questions. The first is about the explanation for the different apparent performance of the breast cancer screening programmes in North West England and the Netherlands. The second involves the robustness of the conclusions that two modifications of the U.K. breast screening programme, extending the age range of the U.K. from upper age 64 to 69 and shortening the screening interval from 3 to 2 years, are roughly equally cost-effective and that the cost per life year gained of both extensions of the programme is not much higher than of the ongoing programme.
Chapter 7 compares the usual method of cervical cancer screening by pap smears with unaided visual examination. The latter is a cheaper screening test that requires less technological input and it is therefore considered as a more feasible alternative for pap smears in developing countries. We show under what model assumptions unaided visual inspection of the cervix is more cost-effective than screening by pap-smears to prevent mortality from cervical cancer.

balancing favourable and unfavourable effects
Cancer screening inevitably leads to unfavourable effects among those being screened. Undergoing screening as such is often an uncomfortable process, undergoing diagnostics for cancer after a positive screening result is also unpleasant and causes grave anxiety in many of those who are affected. There is also the burden from diagnosing cancers earlier or diagnosing cancers that would not have been diagnosed resulting in extra life years with cancer and to more unnecessary cancer therapy. In types of screening that cannot prevent cancer incidence, such as breast cancer screening and prostate cancer screening, the more frequently occurring

unfavourable effects from screening and extra diagnoses are on balance small in comparison with the, though less frequent, stronger health effects from extra incidence and earlier detection of cancer. There are situations where cancer screening may significantly reduce mortality from the disease, but where unfavourable effects are so substantial that screening is not prudent. An example of such a situation is screening for breast cancer at older ages. Even when assuming that the mortality reduction due to screening remains as high at older ages as in the randomised trials, then still the number of life years that can be gained decreases from a certain age. Moreover the probability of finding a breast cancer that would not have been diagnosed before the woman would die from other causes, increases steeply at higher age. Chapter 8 presents estimates of this change in balance with increasing age. Chapter 9 shows that the amount of unfavourable effects relative to the favourable effects (the major reason for an upper age limit for a screening programme) will not be diminished by a longer screening interval.

effectiveness and circumstances

Chapter 10 gives a comprehensive overview of different aspects that are to be taken into account for the evaluation of cancer screening. Most examples are from breast cancer screening. Besides this general framework for screening evaluation, there will still arise compelling questions that do not fit well into this framework. Two examples of such questions are given in chapters 11 and 12. Chapter 11 analyses the question that came up in Denmark where in the discussion about implementing breast cancer screening there was concern for the apparent higher number of cancers detected in the group of women invited for screening in the Malmö trial in comparison to the control group, which might be interpreted as a sign of overdiagnoses. Chapter 12 shows the striking similarity between observed stage distribution at first and repeat breast cancer screenings and discusses difference explanations for this unexpected similarity, since under plausible assumptions one would expect a substantially more favourable stage distribution at repeat screenings.

2. QUANTITATIVE INTERPRETATION OF AGE-SPECIFIC MORTALITY REDUCTIONS FROM THE SWEDISH BREAST CANCER- SCREENING TRIALS

abstract

background: Results from five Swedish randomized trials may provide the most conclusive evidence on the effect of mammographic screening and have been used to forecast the expected reduction in breast cancer mortality in other programs. However, those trials demonstrated different degrees of reduction. The interpretation of observed mortality reduction after long follow-up for women aged 40-49 years at trial entry is both important and controversial.

purpose: We estimated what percentage of the observed mortality reduction for women aged 40-49 years at entry into the five Swedish screening trials might be attributable to screening these women at 50 years of age or older. Moreover, we calculated the most likely percentage mortality reduction for specific screening programs if the Swedish results were generalized and analyzed whether characteristics of each trial might at least partly explain the observed differences in reductions among the trials.

methods: Each Swedish trial was simulated with one underlying computer simulation model (MISCAN-MIcrosimulation SCreening ANalysis) of the natural history of the disease and the performance of screening, taking into account nine important trial characteristics. Improvement in prognosis for screen-detected case patients was estimated with age-specific reduction for all trials and each trial design as a reference.

results: An expected 7% reduction in breast cancer mortality for women aged 40-49 years at trial entry (relative risk [RR] = 0.93) was determined by computer modeling, assuming no improvement in prognosis for cancers that are screen detected before 50 years of age. This result indicates that, of the overall 10% observed reduction (RR = 0.90) in the five Swedish trials analyzed, most (70%) of this reduction might be attributable to screening these women in later rounds after their 50th birthday. Using additional trial information, predictions of breast cancer mortality reduction in women 50 years or older might be 11% larger than previously expected, assuming that high-quality mammographic screening can be achieved in nation-wide programs. For women aged 50-69 years at trial entry, the differences in expected versus observed mortality reduction among the trials are estimated to be relatively small. (Expected mortality reductions range from 24% to 32%.)

conclusions: Results from the Swedish randomized breast cancer-screening trials should be seen as more favorable regarding the effect of mammographic screening in reducing breast cancer mortality for women aged 50-69 years than was estimated earlier. Our analyses also suggest that the improvement in prognosis due to screening for women aged 40-49 years is much smaller than that for women aged 50 years or older. Approximately, 70% of the 10% observed reduction in breast cancer mortality (i.e., 7%) for women aged 40-49 years at trial entry might be attributable to a reduction due to screening these women after they reach age 50.

implications: Detailed screening data for the 40- to 49-year age group of all Swedish trials should be analyzed to specifically estimate the natural history and performance of screening in this age group.

introduction

It is still uncertain whether breast cancer screening for women under 50 years of age is effective in reducing breast cancer mortality (Eckhardt et al. 1994; Fletcher et al. 1993a). After eight rounds of biennial screening in a program generally considered to have reached good quality, the Nijmegen screening project recently showed no positive effect (Peer et al. 1994). According to some investigators, the results from the five Swedish randomized breast cancer-screening trials (Malmö, Kopparberg, Östergötland, Stockholm, and Göteborg) (Chamberlain 1993; Nyström and Larsson 1993; Nyström et al. 1993) can be considered to give the most conclusive evidence on the effect of mammographic screening. The published 10%-13% breast cancer mortality reduction rates for Swedish women under 50 years of age entered in a randomized study may seem encouraging. However, some women in this age group were also screened when they were 50 years old or older. Part of the observed mortality reduction in these women is likely to have been a result of detecting the cancer earlier in later rounds when the women were 50 years old or older, as was seen in the Health Insurance Plan (HIP) trial (Shapiro et al. 1982).

Furthermore, any trial is specific in its design, quality, and background situation. Consequently, different screening trials will result in different breast cancer mortality reductions in the study group as compared with the control group, even for the same age categories. The new and more detailed results from all five Swedish randomized trials, specified per age category (Nyström and Larsson 1993; Nyström et al. 1993), again show that the estimates of reduction vary widely between the trials. It is important to analyze to what extent differences in the characteristics of the five trials are likely to have caused these variations. The earlier publication of the seemingly less favorable results from the randomized

trial conducted in Malmö led to a discussion of the efficacy of mammographic screening (Andersson et al. 1988). The 21% reduction rate achieved for women in the invited age group inclusive of those aged 55-69 years seemed to be in contrast to the earlier published 39% rate in the group aged 50-74 years from the randomized trial in Kopparberg and Östergötland (Tabár et al. 1985). Important characteristics of the trials, such as screening interval, attendance rate, follow-up period, and age groups, should be considered, however (de Koning et al. 1991).

These characteristics of the trials are especially important with regard to screening in younger women. No individual trial has had the power to show a statistically significant mortality reduction in younger women. Much effort is being put into new trials for women under age 50, e.g., the U.K. trial, for which these issues are highly relevant. By using one underlying model that incorporates both the natural history of breast cancer and the performance of mammographic screening, we have analyzed all five trials and have taken into account nine important characteristics within each trial. The different policies for women aged 50 and above or under age 50 at entry are distinguished, as are characteristics in screening practice regarding intervention in the control groups. The goal is to adjust for as many relevant characteristics in screening policy as possible in each trial that may have influenced the outcome and estimate the improvement in prognosis for screen-detected cases. Although each trial may then be unique, the five Swedish trials should help in quantifying the breast cancer mortality reduction expected in other screening programs. Our analysis addressed three questions: 1) Which percentage of the observed mortality reduction for women aged 40-49 years at entry into the trial was likely to have been due to screening these women when they were 50 years or older? 2) What is the extent of breast cancer mortality reduction to be expected for present and. future screening programs, if the Swedish results are generalized to, for example, those from The Netherlands and the United Kingdom? 3) Does the information (more details available) from the five trials at least partly explain the differences in observed mortality reduction and predict differences in efficacy between the five trials?

methods

underlying model of natural history of breast cancer and performance of screening

The computer simulation package MISCAN (MIcrosimulation Screening ANalysis), developed at our institute, was used to evaluate the five screening trials, in which the natural history of the disease, the epidemiology, the design of the screening program, and the performance of screening are incorporated (Habbema et al. 1986; van Oortmarssen et al. 1990b). The natural history of breast cancer is modeled as a progression through a number of stages. The first stage is no breast cancer; women are included in this stage until a transition occurs to one of the preclinical stages when a tumor becomes detectable by screening. There are one ductal carcinoma in situ (DCIS) stage and four invasive stages in the model, according to T-status (T1a, T1b, T1c, and T2+). The duration in the different stages follows an exponential distribution. The transition to the clinically diagnosed stages (with the same subdivision) is governed by the data on incidence and clinical stage distribution. In the case of early detection, women will enter the screen-detected stages (again with the same subdivision). The two end scales of the model are death from breast cancer and death from other causes, based on mortality data. Key parameters in the model of the performance of screening are mean duration of screen-detectable preclinical disease, sensitivity, and improvement in prognosis for screen-detected cancers. Age-specific assumptions with regard to the mean duration of the five preceding screen-detectable preclinical stages of breast cancer and the sensitivity of screening had been validated with the use of all data from the Dutch screening projects (including women aged 35 years or over in the Nijmegen project and women aged 40 years or over in the Utrecht project) covering different periods and screening intervals (de Koning 1993; de Koning et al. 1991). Sensitivity is stage dependent: 40%, 65%, 80%, 90% and 95% (aged 50 years or older) for DCIS, T1a, T1b, T1c, and T2+, respectively. The sensitivity for screening women under age 50 was estimated in the model to be 60% (ages 40-44 years) or 80% (ages 45-49 years) of the sensitivity for screening women older than 50 years. The mean duration of the preclinical screen-detectable period was approximately 1.8 years at age 35 and 6.2 years at age 70. These assumptions resulted in a good fit between model predictions and observed detection rates and interval cancers (both by age, stage, screening round, and interval) in the Dutch screening projects. (Detailed analyses of fit can be obtained from the investigators.) Other epidemiologic parameters (e.g., incidence, stage distribution, and mortality) were based on the Dutch data.

If one applies screening to a population, the shift from diagnosing relatively large clinical cancers toward diagnosing earlier (screen-detected) stages of cancer results in a decrease in breast cancer mortality, as shown in the randomized screening trials. In the model, women with screen-detected cancers can have a reduced risk of dying of breast cancer depending on the cancer size at detection. The degree of this improvement in prognosis after early detection had first been estimated on the basis of

survival differences between women with screen detected cancer and women with cancer diagnosed in the control group of the HIP trial (category-specific estimates based on differences in lymph node metastases), corrected for lead time (Habbema et al. 1986; van Oortmarssen et al. 1990a). An additional amount of improvement had to be (plausibly) assumed to achieve the reduction by the model as observed in more modern trials, such as chose conducted in Malmö, Kopparberg, and Östergötland (de Koning et al. 1991). The number 1 minus the ratio of the risk of dying of screen-detected breast cancer divided by the risk when the cancer had been diagnosed in the absence of screening was estimated to be 0.80, 0.73, 0.51, and 0.35, respectively, for cancers screen detected in stages T1a, T1b, T1c, and T2+. Before the Swedish overview was published, we simulated the Malmö, Kopparberg, and Östergotland trials, using these assumptions and taking into account the differences in design. The simulated reduction in the breast cancer mortality rate (given the natural history, performance of screening, and the epidemiology as described before) was the same as the weighted average reported from these trials for women aged 50-69 years, under the assumption that sensitivity and mean duration of preclinical, screen-detectable disease did not differ strongly from the Dutch situation (Boer et al. 1994; Day et al. 1988; de Koning et al. 1991). Appendices on all the variables taken into account in the model are published in (van Oortmarssen et al. 1990a; van Oortmarssen et al. 1990b).

Swedish trial characteristics and overview of mortality reduction

For the present analysis, each of the five Swedish screening programs was characterized, including intervention for the control group after a specific time (as in the follow-up model). A review of all publications from the five trials, updated with personal information from the trialists, provided nine specific characteristics of all five trials: 1) age distribution at entry, 2) attendance rates for first round in the study group by age, 3) screening interval for study group by age and period, 4) start of intervention in control group, 5) attendance rates for first round in the control group by age, 6) assumed screening interval by age for women from the study and control groups after discontinuation of the trial (start intervention control group), 7) mean duration of follow-up, 8) woman-years per age and study and control category up to December 31, 1989 (Table 2.1). With these characteristics, for all women who entered the trials in either arm or either age group, both the number and timing of screens are determined on the basis of the mean follow-up period, the screening interval, and the attendance pattern in each trial. For example, the women in the study group of the Stockholm trial who had an average follow-up of 8.3 years and an average screening interval of 2.3 years will have had a maximum of four screens. Women in the control group were invited to participate in screening an average of 4.0 years after the start of the trial for the first time and may have had two screens in this analysis (8.3 - 4.0 divided by the 2.3-year interval). The attendance rates in some subsequent rounds are known, but the attendance rates in subsequent rounds of first attenders versus first non-attenders are published for only the Kopparberg, Östergötland, and Stockholm trials and were assumed to be the same in the other trials (characteristic 9). If no information was available, we assumed that, after the first screening of the control group, women had been invited to participate in screening up to December 31, 1989, according to actual age and according to the initial screening policies. The screening policy and, therefore, the number and timing of screens after age 50 for women who entered the studies at 40-49 years of age are certainly different in the intervention and control arms of the Kopparberg and Östergötland trials, since each trial had a different screening interval according to age at entry. Information was not available for all trials on the age-specific percentages of women in the control group who had mammography for non-symptomatic reasons before being invited to participate in the screening. We have not assumed any difference in sensitivity with regard to one-view or two view mammography.

The underlying model of the natural history of disease, the age-specific sensitivity and the age-specific improvement in prognosis were assumed to be the same for all the trials, and the expected case fatality (and breast cancer deaths differences between the study and control groups) was compared with the observed, for all trials together, taking into account the different designs of the five trials. The observed number of breast cancer deaths in each trial by age group at entry into the trial was based on the recent combined (independent) analysis of all causes of death among breast cancer patients in the trials, using the so-called follow-up model (Nyström and Larsson 1993; Nyström et al. 1993). By having detailed information on the number of deaths, screening policies, and designs for younger and older women, we could estimate the improvement in prognosis for screen-detected cases specifically per age category in the present analysis based on the recent over-view. Since we had earlier estimates on improvement in prognosis based on the Malmö, Kopparberg, and Östergötland trials, we could compare these estimates with the new estimates. Finally, the expected consequences of the U.K. and Dutch screening policy were recalculated on the basis of the new information.

Table 2.1 Characteristics of the five Swedish randomized breast cancer-screening trials used in predicting effect on breast cancer mortality reduction rates with one model of natural history of breast cancer and screening quality

Characteristic	Trial				
	Malmö	Kopparberg	Östergötland	Stockholm	Göteborg
Age at entry, y	45-69	40-74	40-74	40-64	40-59
Attendance of women invited for mammographic screening, first round, %					
<50	79	94	93	81	85
50-59	78	93	90	80	83
60-69	68	91	85	81	--
≥ 70	--	81	76	--	--
Screening interval, y (age range)	1.75*	2,0 (40-49) 2.8 (≥50)‡	2.5,2.0† (40-49) 2.5, 2.8† (≥50)‡	2.3	1.5
Attenders'(attending previous round) attendance, %					
< 50	85	94	93	95	87
50-59	85	95	91	95	88
60-69	85	91	85	95	--
≥ 70	--	80	72	--	--
Nonattenders' (not attending previous round) attendance, %					
< 50	30	50	29	16	40
50-59	30	43	30	16	28
60-69	30	33	27	16	--
≥ 70	--	37	25	--	--
Intervention control group, y after randomization (age range)	--	6.8	7.0	4.0	7.0 (40-49) 5.0 (50-59)
Attendence rate, %, first-round control group					
< 50	--	90	90	77	67
50-59	--	87	87	77	78
60-69	--	80	80	79	--
Assumed screening interval, y, control group /study group after stopping trials	1.75	2.0/2.8§	1.25/2.25§	2.3	1.5
Mean follow-up, y	11.8	10.2	9.7	8.3	6.2

* Average, depending on breast density. † First round, subsequent rounds. ‡ Only two screens ≥ 70. § According to actual age.

statistical analysis

Adaptation of the model was tested by using the sum of the squared differences between the observed and expected numbers of breast cancer deaths in each of the five trials divided by the estimated variance. The expected number of deaths and variance were derived by assuming the total number of deaths as observed and a binomial distribution over the study and control groups. The outcomes are chi-squared distributions with four degrees of freedom. We used the Mantel-Haenszel method to calculate the average of the relative risks (RRs) modeled for the different trials. The model was fitted to the observed average values of 0.70 (ages 50-69 years) and 0.90 (ages 40-49 years). The precision of these values was equal to what was observed.

results

Table 2.2 shows the observed number of breast cancer deaths and RRs in all study and control groups combined per age category as well as the expected numbers and RRs obtained from the model. In all variants, we used the same underlying model of the natural history of disease and performance of screening for each of the five trials (i.e., sensitivity and mean duration of preclinical, screen-detectable disease by age and stage). The simulation of the specific Swedish trial designs then leads to different numbers of screen-detected cancers in the different stages, with a consequent reduced risk of dying of breast cancer. Also, one identical parameter for improvement in prognosis was assumed for all trials. First, this stage- and age-specific parameter was estimated to make the results of the model for all

Table 2.2 Woman-years and observed numbers of breast cancer deaths in study and control groups and RRs for all five Swedish trials combined, per age category, compared with expected ones with the model having different assumptions on improvement in prognosis for screen-detected cases for women aged 40-49 years

Woman-years* (x1000)		Observed breast cancer deaths* (all 5 trials combined)			Expected breast cancer deaths (by computer model)		
Study group	Control group	Study group	Control group	RR	Study group	Control group	RR†
(a) Women aged 50-69 years entering trial; assuming an observed overall 30% reduction (RR = 0.70) in the risk of women aged 50-69 years dying of breast cancer for all trials, and fitting this same reduction to the model (RR = 0.70) for consistency.							
911	725	281	312	0.70‡	275.3	317.7	0.70§
(b) Women aged 40-49 years entering trial; assuming an improvement in breast cancer prognosis equal to the one estimated for women aged 50 years or above as in a.							
428	350	84	75	0.90*, ||	79.5	79.5	0.77¶
(c) *) Women aged 40-49 years entering trial; assuming no improvement in prognosis for cancers that are screen detected before 50 years of age and improvement in prognosis equal to the one estimated for women aged 50 years or above as in a.*							
428	350	84	75	0.90*, ||	86.7	72.3	0.93#
(d) Women aged 40-49 years entering trial; assuming improvement in prognosis for women aged 40-49 years with screen-detected cancer to be much lower than that estimated for women with screen-detected cancer at ages 50 years and above and the fitting of expected RR of the model to precisely that of the observed RR.							
428	350	84	75	0.90*, ||	85.3	73.7	0.90**

*Observed number of breast cancer deaths in each trial by age at entry based on recent combined. independent analyses of all causes of death in die trials using the follow-up model and trial end point of December 31, 1989 (*45*); women aged 70 years or older at trial entry were excluded and estimated from numbers of women at entry.

†Calculated by the Mantel-Haenszel method.

‡Average of published RRs for all five trials (0.72 for age group 50-59 years, 0.69 for age group 60-69 (0.72 + 0.69)/2 = 0.70).

§RR estimated to make the result of the model for all trials consistent with the most current observed overall 30% reduction (RR = 0.70) for ages 50-69 years. RR was estimated using the same stage and age-specific parameter for improvement in breast cancer prognosis for all five trials.

||Observed reduction in the risk of dying of breast cancer in women aged 40-49 years for all trials combined was 10% (RR = 0.90), one third of that for the age group 50 years or older (RR = 0.70:30% reduced risk of dying of breast cancer).

¶Expected overall 23% reduction rate in breast cancer mortality (RR = 0.77) for all women aged 40-49 at trial entry: 1) assuming the improvement in prognosis to be eaqual to the improvement estimated for women aged 50-69 years and 2) taking into account the characteristics and the follow-up periods of all trials.

#Expected 7% reduction rate in breast cancer mortality (RR = 0.93) for women aged 40-49 years at the start of the trials where there is no benefit in the model for women whose cancer was detected by screening before age 50 (compared with no screening). This 7% benefit must be a model-derived result for women whose cancer was detected at age 50 or over, where a reduction in risk of dying of breast cancer was assumed. This expected 7% reduction rate (RR = 0.93) indicates (that most of the 10% observed reduction rate (RR = 0.90), seven of 10 (70%), might be attributed to screening these women in the later rounds when they were already 50 years or older.

**Expected overall RR adjusted to Fit the observed mortality reduction for all trials combined (RR = 0.90). Improvement in prognosis for women aged:140-49 years with screen-detected cancer was assumed to be much lower than that estimated for women aged 50 years or older with screen-detected cancer.

trials together consistent with the newest observed overall 30% reduction rate for ages 50-69 years (line a). Table 2.2 also shows the results from the trials and the model for women aged 40-49 years (or 45-49 years) at the time of random assignment (lines b-d). The observed reduction rate in the group aged 40-49 years for all trials combined was 10%, a third of that in the age group 50 years or older. All screen-detected cases resulting from the design and the underlying model have been given the same improvement in prognosis (line b), equal to the one estimated for women above 50 years of age (as in line a). With that assumption, we would have expected an overall 23% reduction in breast cancer

mortality (RR = 0.77) for all women aged 40-49 years at trial entry, given the characteristics and the follow-up periods of all trials (line b) Assuming no improvement in prognosis for cancers that are screen detected before 50 years of age and the same improvement as in line a for cancers that are screen detected at age 50 years or more, a 7% mortality reduction between the invited groups and the control groups would have been expected at the end of follow-up of these five trials (line c). In other words, in a so-called pessimistic variant in which there is no benefit in the model for women whose cancer was detected by screening before age 50 (compared with no screening), we expected in this analysis that the five trials would have shown a 7% reduction on December 31, 1989, for women aged 40-49 years at the start of the trials. This expected 7% benefit, therefore, must have been derived from the model for women whose cancer was detected at age 50 years or more, where we did assume a reduction in risk of dying of breast cancer. This result shows that most (seven of 10 [70%]) of the observed reduction might be attributed to screening these women in later rounds when they were already 50 years old or older. To simulate the observed mortality reduction for all trials combined (RR = 0.90), the improvement in prognosis for women with screen-detected cancer in this age group had to be assumed to be much lower than that estimated for women above 50 years of age with screen-detected cancer; in line d, the parameter has been adjusted to fit the expected overall RR in this age category precisely to the observed RR. Earlier data published in 1988 and 1989 had shown an RR of 1.15 (Malmö, ages 45-54), 0.79 (Malmö, ages 55-69), and 0.62 (Kopparberg and Östergötland, ages 50-69) between study and control groups (Andersson et al. 1988; Tabár et al. 1989). In earlier analyses when advising about the possible introduction of screening in The Netherlands (published in 1991), we had found no evidence for a difference in efficacy between these trials on the basis of screening policies. If adjustments were made for interval, attendance, follow-up period, and age groups, these characteristics were expected to compensate for each other (de Koning et al. 1991). The then resulting 32% weighted (on size of confidence intervals) average reduction rate in breast cancer mortality for women aged 50-69 years who were invited to participate in the Malmö, Kopparberg, and Östergötland trials seems, however, better than the most recent results published in 1993 from all five Swedish trials, with 28% in the age group 50-59 years and 31% in the age group 60-69 years (average, 29.5%) used for this analysis (Nyström et al. 1993). In the present analysis, however, the Göteborg and Stockholm trials are included with less broad age ranges, intermediate attendance rates, and other screening intervals. Furthermore, more detailed information with regard to the characteristics of the trials has become available. In fact, we now would have expected a smaller overall reduction for all five trials with the model of 26.5% (weighted on trial sizes) if the old estimate (de Koning et al. 1991) on improvement in prognosis for screen-detected cases would still hold. The newest published results from all trials are, therefore, compatible with an 1-1% (29.5/26.5 times) larger improvement in prognosis for screen-detected case patients aged 50-69 years at randomization than previously expected. The predicted breast cancer mortality reduction (not trial level) in the total population in The Netherlands with mammographic screening once every 2 years for women aged 51-69 years (Dutch nation-wide policy) or once every 3 years for women aged 51.5-63.5 years (U.K. nation-wide policy) can now be re-estimated, given the performance of screening in The Netherlands and this interpretation from the Swedish trials. The estimated improvement in prognosis on the basis of the five Swedish trials is used for the analysis of mortality reduction. As a consequence of our analysis, the predictions on reduction can be adjusted in a more favorable direction, based on the assumption that the high quality of screening can be achieved in such nation-wide programs (de Koning et al. 1995b). In The Netherlands, a 17% reduction in the annual total female breast cancer mortality rate seems realistic (meaning >800 breast cancer deaths prevented per year). The expected reduction rate of 11 % for the other practices (15% in the group invited to participate in the screening) is likely to be somewhat higher in the United Kingdom because of the presumably worse clinical stage distribution compared with that seen in The Netherlands. For all trials together, we were able to make a good fit (agreement) between the observed reduction and the model. Table 2.3 shows the expected RRs for each trial and compares them with the observed RRs. In general, for women aged 50-69 years at trial entry, the differences in the expected mortality reduction rate between the trials are estimated to be relatively small (range, 24%-32% reduction), considering the different trial designs. It can be seen that the mortality reduction in this age group is expected to be the smallest in the Stockholm trial because of its specific characteristics. The Malmö and Göteborg trials' (with the shortest screening intervals) possibly could have been expected to lower the mortality relatively more than the other trials. In reality, the observed reductions varied more widely (9%-39%). Both the Kopparberg and the Stockholm trials have produced better RRs than expected. On the basis of our analysis, we expected a better RR for the Göteborg trial and the Malmö trial than was observed. The numbers, however, clearly were small, and the variance found in outcome was not larger

Table 2.3 Expected RRs in each trial per age category if improvement in prognosis is based on all trial results (top line), but individual trial characteristics as in Table 1 are taken into account, compared with observed RR (each trial assumed to have the same improvement in prognosis for screen-detected cases)

	50-69 y				40-49 y			
	Women years* (x 1000)			Observed RR* (deaths per study group)	Women years* (x 1000)			Observed RR* (deaths per study group)
Trial	Study group	Control group	Expected RR		Study group	Control group	Expected RR	
All	911	725	0.70†	0.70‡(281)	428	350	0.90†	0.90*(84)
Malmö	193	193	0.68	0.86 (79)	46	47	0.82	0.51 (8)
Kopparberg	249	115	0.70	0.61 (79)	107	56	0.90	0.76 (26)
Östergötland	224	213	0.71	0.69 (69)	104	106	0.90	1.29 (24)
Stockholm	180	100	0.76	0.65 (33)	107	64	0.95	0.99 (20)
Göteborg	65	104	0.68	0.91 (21)	64	77	0.92	0.72 (6)

*Observed number of breast cancer deaths in each trial by age on recent combined, independent analyses of all causes of death in the trials using the follow-up model and trial end point of December 31. 1989 (4,5): women aged 70 or older at trial entry were excluded.
†Calculated by the Mantel-Haenszel method.
‡Average of published RRs for all five trials (0.72 for age group 50-59: 0.69 for age group 60-69: [0.72 + 0.69]/2 = 0.70).

than that expected from trials with the same design (chi-square = 5.05). Random fluctuation in the trial results is the most important explanation for the remaining discrepancies. Still, one can argue that differences in the nine important characteristics of the screening trials that were included in our analysis do not satisfactorily explain the observed differences in breast cancer mortality reduction between the five trials.

For women aged under 50 years at trial entry, Table 2.3 shows that the Malmö trial (which, for example, has a short screening interval, no official intervention in the control group, and the longest duration of the trial) would be expected to result in the largest breast cancer mortality reduction rate (RR = 0,82). The two most recently started trials, Göteborg and Stockholm, are expected to show only small reductions (RR = 0.92 and RR = 0.95, respectively). It is clear that the variation between the trials in the reductions (expected) is larger for this age group (range, 5%-18% reduction), again given the different designs and/or follow-up periods. The comparison between expected and observed results for this age group in each trial is, of course, strongly hampered by the small number of women invited to participate and the number of deaths expected (chi-square = 3.20). This situation would make a favorable conclusion about the Malmö trial for this age group still speculative. If there are no important details found from the Östergötland trial, other than those taken into account, that differ strongly from the other four trials, these are probably the best estimates on mortality reduction, including a strong negative effect from this specific trial. The Kopparberg trial has better RRs than expected for both young and older women.

discussion

This analysis shows that the newest results from the Swedish randomized breast cancer-screening trials should be seen as more favorable with regard to the effect of beast cancer screening for women aged 50-69 years than earlier estimated (only from the reports from Malmö, Kopparberg, and Östergötland). Although the weighted average observed reduction reported presently is smaller than the earlier published average, our analysis shows that specific characteristics of all five trials are responsible for this. The longer follow-up, the additional details about the programs, and especially the information about dilution of the effect due to intervention in the control groups have been important. It is difficult to say whether the blind and uniform ascertainment of the deaths from breast cancer by an independent panel has influenced results, but for women aged 70 years or older and for women under age 50 years in the Östergötland trial, the absolute numbers now classified do differ strongly from the earlier published numbers (Nyström and Larsson 1993; Nyström et al. 1993; Tabár et al. 1989). Our analysis also led to the assumption that the improvement in prognosis due to screening for women aged 40-49 years is much smaller than for women 50 years old or older. About 70% of the reduction observed in the trials' results for women aged 40-49 years at entry

into the trial might be attributed to a reduction due to screening these women when they were 50 years or older. This fact should be borne in mind in the expectations for any trial on women under age 50 years. These results are important for a correct interpretation of the possible achievement of screening in present or future programs (Beemsterboer et al. 1994). Although any trial or program is unique, we think that a 25% reduction rate for the invited group aged 50-64 years in the U.K. screening program as estimated by other investigators (Vessey 1994) seems high, even if this overview analysis would have been based on the Kopparberg and Östergötland trials only. We believe that the present method provides a good (and maybe the last) opportunity for predictions to be made on mortality reduction at a nation-wide level. The conclusion of the trials and the screening, of the control groups will influence the differences in breast cancer death rates in both groups with longer follow-up. Also, to interpret the observed mortality differences in the future, analyses including the effect of intervention in the control groups will be required. One should be cautious in interpreting the differences in this analysis between the observed and expected RRs in each trial.

Without modeling, clearly, the numbers are so small that the variance in outcome found is not larger than expected from trials that have the same design. Random fluctuation in the trial results is the most important explanation for the discrepancies, but two other possible explanations are important. On the one hand, certainly a number of details from the trials or the baseline situation in the populations studied have not been available to us: regional epidemiology (population structure, incidence by age, stage distribution and treatment, and survival), details about the influence of the type of random assignment, and especially details about the situation after the first screening of the control group. Still, given the important details taken into account for each trial, this information is unlikely to strongly influence the present results and conclusions concerning women aged 50-69 years. On the other hand, other factors might have influenced the outcome of screening. Further research should be initiated to quantify the quality of screening, especially in younger women.

Such detailed information is not available in the literature on all five trials (Tabár et al. 1995). Although it seems appropriate to base this analysis with respect to the natural history of breast cancer and age-specific sensitivity partly on Dutch data, where there are no data available to assume a reasonable difference between both countries (Day et al. 1988; de Koning 1993), detailed information on detection rates, interval cancers, stage distributions, and the background situation in each Swedish trial are needed. It would then be possible to estimate whether our assumptions with regard to either natural history or sensitivity might have to be adjusted. The especially interesting question is whether it is indeed correct to attribute most of the published mortality reduction for women less than 50 years of age who were invited to screening above this age group. A standard meta-analysis does not account for the underlying parameters causing a possibly different RR in a particular screening situation.

A further analysis is needed with detailed Swedish screening data on women aged 40-49 years. Analysis of the other randomized trials could be performed as shown in this article. Although such an analysis as ours could never replace actual data from the different randomized, controlled trials, it would be very meaningful. With longer follow-up, the Swedish trials will likely show a higher or statistically significant breast cancer mortality reduction for the women aged 40-49 years at trial entry, but the analysis shows it will be crucial to get a reasonable estimate with regard to the amount of reduction achieved for these women on the basis of screening in later rounds.

additional discussion

screening for breast cancer under the age of fifty

When the preceding article was published in the Journal of the National Cancer Institute, it got accompanied by two editorials and later there was a letter to the editor about the article.(Forrest and Alexander 1995; Smith 1995; Tabár 1996) Though only a minor part of the article was about screening under age 50, all three comments concentrated on that issue.

At the time of publication of the article, the overview of the Swedish randomised trials into breast cancer screening showed a substantially, though not statistically significant, lower mortality reduction for women under age 50 at entry compared to women aged 50-69 at entry.(Nyström et al. 1993) Apart from that observation there was also the outcome from the Canadian NBSS that was quite unfavourable for screening women in their forties.(Miller et al. 1992) A problem in the interpretation of the published trial outcomes is that women who are classified as under age 50 at entry to the trial, may have benefited from screening at ages of 50 and higher. Our analysis assumed that that the efficacy of screening suddenly changes at age 50 and that efficacy by age is constant within each of the age ranges of under 50 and over 50. Using these assumptions we showed that the observed screening benefit of women under age 50 at entry to the trial could be attributed for the largest part to screening from age 50 onward. We did not conclude that screening under age 50 was worthless, but that this called for further investigation. The mentioned letter to the editor was a reply to this

question and showed that in the Two Counties Study, there was evidence against screenings applied over age 50 being more effective than those applied under age 50 among the age group of women under age 50 at entry.(Tabár 1996) Together with the relatively low observed efficacy of screening under age 50 at entry this implies that the screenings applied over age 50 in that age group were remarkably little effective. But because the numbers involved are so small, it is not prudent to draw any conclusions from that.
In the mean time outcomes of the Swedish trials for a longer follow-up have been published, showing a much smaller difference in mortality reduction by age at entry. But the Canadian NBSS study in women under age 50 still shows an unfavourable effect on mortality from screening. The issue is continuously under vigorous discussion (see for instance (Baines and Miller 1997; Chang et al. 1997; Cox 1997; Feig and Hendrick 1997; Fletcher 1997; Harris 1997; Kopans 1997c)). Some proponents of screening younger women appear to be biased in their arguments. For instance the NBSS trial has been under severe attack, while other trials are also liable to more or less similar criticism.(Baines 1994; Glasziou and Irwig 1997; Gøtzsche and Olsen 2000; Kopans 1997a; Kopans 1997b)

current evidence for efficacy of breast cancer screening under age 50

Except for the Canadian NBSS trial, none of the trials were designed to show a significant mortality reduction for women younger than 50 years. They were designed to show an effect in the total trial population. The single fact that they did not show a statistically significant mortality reduction for separate age groups should not be interpreted as a lack of efficacy. However, it is possible that after the start of a trial new ideas develop on questions that need to be answered based on empirical evidence. Clearly, the question whether screening under age 50 is efficacious, is such a question. During the course of the HIP trial it was observed that the mortality reduction was lower among women under age 50 at entry than for the other participants.(Shapiro et al. 1974) This was an important spark off for the discussion, with as fuel, the notion of important biological changes occurring around the age of 50, or rather, the menopause. After longer follow up of the HIP trial, the difference in mortality reduction almost disappeared.(Shapiro et al. 1985) And nobody has substantiated why any particular biological change that occurs around the age of 50 would cause a large difference in efficacy of screening. Dense breasts and faster growing tumours indeed tend to cause a somewhat lower efficacy, but the screening of women under age 50 still detects a very substantial proportion of the cancers that surface in the study and at a practically equally favourable stage as in older women. Therefore, dense breasts and faster growing tumours alone cannot explain a large difference in mortality reduction among women over and under age 50.
The ongoing discussion resulted in two trials that focussing on screening women under age 50. The first is the Canadian NBSS that started in 1980 and shows a (statistically non-significant) higher mortality from breast cancer in the screening group than in the control group.(Miller et al. 1992; Miller et al. 1997) After this result, of course there was sufficient reason for concern about the efficacy of screening under age 50. Therefore a second trial was initiated in the U.K., which is so recent that no results are available yet.
Given the results from the collective randomised trials on breast cancer screening and the discussion about biological influences, it appears that screening under age 50 can reduce breast cancer mortality. But there remains a compelling question why the Canadian trial that was designed to show an effect among women in that age group does not show a reduction in breast cancer mortality.
Criticism on the trial has been convincingly countered.(Baines 1994) What appears to be lacking is a detailed comparison of all intermediate outcomes from this trial with those from other studies. Such a comparison may be facilitated by using a model such as Miscan that can show the influence of local circumstances on the observed results. Insight in the circumstances that determine whether one trial shows a mortality reduction, while another does not, is crucial for any future breast screening. As also chapter 12 shows, our understanding of the natural history of breast cancer in relation to screening is still rather incomplete. A more detailed study of the results from the Canadian and other trials may enhance our understanding of breast cancer screening that in turn may indicate possibilities for improving screening efficacy.

choosing the lower age limit of mass screening for breast cancer

Sufficient evidence for mortality reduction due to screening does not imply that the balance of all favourable and unfavourable effects (including costs) is sufficiently positive. And the best estimate of the balance may be sufficiently positive, but that estimate may also be so uncertain that a decision based on it would not be called 'evidence based'. Both issues have been under discussion, but the emphasis was thus far on the sufficiency of the evidence. Now that the evidence in favour of screening women at ages under 50 has been increasing over the past several years, there is more need for good estimates on the balance of favourable and unfavourable health effects as well as that of effectiveness and costs. What has been published so

far on the issue of balancing can at best be called 'tentative'.(Boer et al. 1995b; Salzmann et al. 1997) Although very different estimates of the balance between favourable and unfavourable effects have been given, no one concluded that there are inevitable unfavourable effects that would preclude screening of women in their forties. Also on cost-effectiveness estimates differed widely, and here using one or the other estimate would make the difference between deciding for or against screening women in their forties. At some points the estimates differ on very fundamental issues. For instance, our research group maintains that in general the law of diminishing returns applies to screening frequency, thus more frequent screening is less cost-effective. Other research groups estimate a better cost-effectiveness for yearly screening than for two-yearly screening.(Chen et al. 1997; Michaelson et al. 1999) There is clearly a need to gather all available information in order to make a more precise estimate of the balance of effectiveness and costs of breast cancer screening of women under the age of 50, and need for detailed comparison of the different models that have been used to estimate favourable and unfavourable effects, and costs.

There appears to be general agreement only on some issues. Breast cancer is also under the age of 50 an important health problem. Though incidence and mortality from the disease are substantially lower than over the age of 50, the number of life years lost per death from the disease is higher. And due to faster growth of tumours at lower ages, and possibly also lower sensitivity of the screening test, reaching the same percentage of mortality reduction needs more frequent screening. That limited agreement is regretfully not a very good basis for decision making.

furthering the method of estimating improvement of prognosis due to screening

The study described in this chapter only accounted for a limited number of aspects that influence screening efficacy. The model was only adjusted for the aspects of organisation of the screening project and attendance of the invited women. When the study was performed, we assumed that the epidemiology and screening performance in the Swedish trial were similar to those in the Netherlands. Later, a more detailed comparison of the Dutch experience with the Two Counties study showed that there are substantial differences.(Fracheboud et al. 1997) Particularly, the Two Counties study appeared to pick up about twice as many small (up to 1 cm) invasive tumours than the Dutch screening programme. In order for the model to predict the same mortality reduction, but with a kind of screening that finds more, particularly smaller, cancers, the probability for cure due to detection at screening should be lower. Thus, with our current understanding, the estimates presented in the article overestimate the efficacy of screening, therefore the prognoses for effectiveness of screening that were made with this estimate are too optimistic. And more importantly, there is reason to assume that the Dutch breast cancer screening programme can possibly perform better. Currently a study for optimisation of the Dutch programme is on its way. It would be useful to repeat our study with taking screening performance into account so that a more accurate estimate of improvement of prognosis due to detection at screening will be available.

3. AN INVESTIGATION OF DESIGN AND BIAS ISSUES IN CASE-CONTROL STUDIES OF CANCER SCREENING USING MICROSIMULATION

abstract

Using a microsimulation approach, the authors examine design and bias issues in case-control studies of cancer screening. Specifically, they look at the impact on the odds ratio of the way in which exposure to screening is defined, the type of age matching, the time scale used, and the criteria used for determining control eligibility. The results show that defining exposure as "ever/never" screened produces, as expected, a serious bias in favor of screening. Defining exposure as being screened no later than the time the case's cancer is diagnosed has a serious bias against screening. An alternative exposure definition-screening can occur no later than the time the case would have been clinically diagnosed-eliminates the bias against screening. Further, the results show that the type of age matching and the time scale used can produce a bias against screening, and that this bias can be quite strong when case-control studies are done in populations with a periodic screening program which is the only source of screening. Finally, control eligibility criteria have small effect.

introduction

Considerable interest has been generated regarding the methodology of cancer screening case-control studies, with the main focus on the design (Aristizabal et al. 1984; Celentano et al. 1988; Clarke and Anderson 1979; Collette et al. 1984; Ebeling and Nischan 1987; MacGregor et al. 1985; Oshima et al. 1986; Palli et al. 1986; Selby et al. 1992; Verbeek et al. 1984) and more recently on the relationship between theory and practice (Cronin et al. 1998). This interest is likely due to the increasing use of the matched case-control design as an alternative to a randomized controlled trial to evaluate cancer screening (Berrino et al. 1984; Friedman et al. 1995; Gill and Horwitz 1995; Hosek et al. 1996; Morrison 1982; Moss 1991; Sasco et al. 1986; Weiss 1983; Weiss and Lazovich 1996; Weiss et al. 1992), whether the goal is to detect and treat early cancers (e.g., breast cancer screening) or precancerous lesions (e.g., cervical cancer screening).

Many questions remain regarding the appropriate design of such studies. In this paper, we use a microsimulation approach to examine how design choices affect mortality odds ratios from case-control studies of screening for breast cancer. We limit our considerations to a setting in which screening is available only through an organized periodic screening program. Microsimulation is a computer-based technique for creating hypothetical individuals whose life histories are created and maintained in the database and thus constitute a population. Incidence and survival from breast cancer and deaths from other causes are programmed into a complex stochastic model using parameters from published data. The computer program ages each individual, updates their disease status according to a complex set of algorithms, and changes their survival time after detection in a hypothetical screening program offered to the hypothetical individuals. Microsimulation with its focus on individuals has been used in a number of areas. In the public health care area the uses include the study of the transmission of infectious diseases (Habbema et al. 1996) and the study of the costs and benefits of cancer screening programs (Habbema et al. 1987; Habbema et al. 1985; van Oortmarssen et al. 1990b). Microsimulation differs from traditional statistical simulations and from "Monte Carlo" simulations. Statistical simulations usually are concerned with small sample properties of equations representing analytic formula rather than properties of hypothetical populations. Monte Carlo simulations tend to focus on the behavior of subgroups within a hypothetical population rather than on the behavior of the individuals comprising the hypothetical population. Nevertheless, in each case simulation studies involve repeated runs and considerable flexibility in changing parameter values and inputs to reveal how the underlying model performs under different conditions

In particular, we use microsimulation to examine whether various design choices for matched case-control studies of screening result in biased odds ratios. These odds ratios estimate the ratio of the mortality rate of those screened to what it would have been had they not been screened, (i.e., the efficacy of screening). We examine the effect of (i) three different measures of exposure, (ii) the way in which controls are age matched, (iii) the choice of the time scale, and (iv) the eligibility criteria used to determine the controls for each case. A microsimulation computer program, MISCAN (MIcrosimulation Screening ANalysis) (Habbema et al. 1985), is used to generate individual life histories in the presence and absence of a breast cancer screening program. Case-control studies are performed using the MISCAN-generated population with screening offered. Definitions of screening exposure, the type of age matching, the choice of time scale, and the control eligibility criteria are varied. Moreover, two screening scenarios are modeled, one with no benefit from screening and the other with a benefit. How well the odds ratio estimates the efficacy of screening under these

varying conditions is assessed by comparing the expected odds ratio with the average efficacy observed.

A stochastic microsimulation approach is used because a closed solution for an analytic model that reasonably reflects the complexity of population screening is not tractable. As noted above the microsimulation approach generates populations of individual life histories within which case-control studies can be performed. The breast cancer MISCAN model was used in this study because it was well documented and had been validated for breast cancer screening(Habbema et al. 1987; Habbema et al. 1985; van Oortmarssen et al. 1990b). Our general approach to the microsimulation and analysis is presented below. Details of the simulation model and the analysis are also provided.

materials and methods

general approach

The output from one MISCAN "run" provides: 1) a population of 50,000 life histories when screening is not available and is not offered and 2) the life histories of these same 50,000 individuals with a "hypothetical" organized screening program in which screening is offered through the program and is not otherwise available. Using the latter population, a matched case-control study is performed. An odds ratio is calculated providing an estimate of the efficacy of screening. Using the two populations, the true impact of screening is directly assessed by comparing the outcomes with and without screening. Odds ratio calculations use all eligible controls for each case. This eliminates sampling variation when a n-to-1 matched design is used, i.e., the variation in the odds ratio due to each case's n controls being randomly sampled from the case's potential set of controls. To investigate the effect on the odds ratio of the definition of screening exposure, the type of age matching, the choice of time scale, and the control eligibility criteria, case-control studies were performed for various combinations of these factors. To estimate the odds ratio expected for a specific combination of factors, the odds ratio was calculated in each of 100 populations of 50,000 life histories generated in 100 MISCAN runs. The average of these 100 odds ratios was used to estimate the odds ratio expected for a specific combination of factors. One hundred runs were used to insure adequate precision in estimating the odds ratios. The standard error of the estimated odds ratio was calculated to assess precision.

The true efficacy of screening expected for a specific combination of factors was estimated by averaging the efficacy obtained in each of the 100 runs. The efficacy for a particular run was calculated directly using the individual life histories with and without screening.

Two screening scenarios were used. The first reflects the situation in which there is no benefit from screening, i.e., women with breast cancer die at the same time that they would have if screening had not been offered. Put another way, there is no reduction in the number of breast cancer deaths among those screen-detected. For this "no benefit" scenario, the efficacy is 1.00. The second scenario reflects a situation in which screening confers a benefit. Here, it is assumed that there is a 50 percent reduction in the number of breast cancer deaths among those screen-detected. The resultant efficacy is less than 1.00 and it is estimated as indicated above.

screening program simulated

The MISCAN model simulates individual life histories taking into account three general areas of assumptions: demography, natural history and screening effects (Habbema et al. 1985; van Oortmarssen et al. 1990b).

Demography. The population created by MISCAN has an age structure and mortality from causes other than breast cancer based on the Dutch female population.

Natural History. The natural history of breast cancer is modeled as a progression through several states. The first state is "no breast cancer." Women reside in that state until a transition occurs to one of the preclinical states when the tumor becomes detectable by screening. In the model used for this study, there are four preclinical disease states, DCIS (Ductal Carcinoma In Situ) and three invasive states according to tumor size (< 1 cm, 1-2 cm and 2 cm), and four corresponding clinical states. The duration in the different preclinical states follows an exponential distribution. The mean duration of the preclinical screen-detectable period increases from approximately 2.7 years at age 50 to 6.2 years at age 70. Incidence and clinical stage distribution data govern the transition to the clinically diagnosed states. The incidence of breast cancer in the model follows that in the Dutch National Hospital Registry(SIG). After a diagnosis of breast cancer, the survival period depends on the disease state and age at time of diagnosis. Breast cancer mortality resulting from incidence and survival closely follows Dutch breast cancer mortality.

Screening. Within the MISCAN model, various assumptions about screening are made regarding the timing of screenings, the detection of cancer by screening, and in particular, the impact of screening. In this study each woman without an earlier diagnosis of breast cancer is invited for screening at each of her birthdays from age 50 through 70 during the study period from 1 January 1990 through 31 December 2004. The population of the case-control studies consists of those women who are invited at least once. Screening attendance is modeled on

Dutch experience (Collette et al. 1984; Verbeek et al. 1984). The probability of attending a first screening is 0.75. When a woman is subsequently invited to attend a later round of screening, attendance depends on her behavior at the immediately preceding invitation: a woman who attended the previous screening has a probability of 0.85 of attending her next scheduled screening. For those who did not attend the previous screening the probability is 0.20. A screening examination consists of two tests: mammography and clinical breast examination (CBE). Given the preclinical disease state at the time of screening, there is an assumed sensitivity of detecting the tumor in that state. Sensitivity of mammography for DCIS is 0.40, for invasive tumors of < 1 cm it is 0.70, and for larger tumors it is 0.95. Sensitivity of CBE for DCIS and for invasive tumors of <1 cm is 0.00, for tumors 1-2 cm it is 0.50 and for tumors 2 cm it is 0.70 (van Oortmarssen et al. 1990b). The effect of detecting a cancer at screening has been simplified for easier interpretation of the outcomes of case-control studies: there is either no mortality benefit from the screening (i.e., all breast cancers follow their natural history as if screening had not been offered) or there is a benefit with a 50 percent reduction in the number of breast cancer deaths among the screen-detected cancers (i.e., 50 percent of those detected at screening with breast cancer who would die from breast cancer without screening do not die of breast cancer, but die at a later time from another cause, as governed by the demographics of the population model).

analysis program

Case-Control Study Odds Ratio. Each MISCAN-generated population of 50,000 individual life histories with screening being offered is converted into matched case-control files with the cases defined as those diagnosed with breast cancer and dying from breast cancer in the 15 year study period, 1 January 1990 through 31 December 2004.

Using the cases and eligible controls, the odds ratio is calculated using the SAS conditional logistic model package, PROC PHREG (SAS 1991).

For each MISCAN-generated population, case-control studies are done for all combinations of exposure measures, type of age matching, choice of time scale, and control eligibility criteria. Categories for each factor are:

1. Exposure to screening:
   A. one or more screens regardless of when they occur (n.b., It is recognized that this is not an appropriate measure of exposure since once a breast cancer is detected, no further screening can occur for that individual and therefore, controls can have a greater opportunity to be screened. This measure is used for comparative purposes and to illustrate that our methodology does produce the expected bias.),
   B. one or more screens but only those that occur before or at the time the case's cancer is diagnosed,
   C. one or more screens but only those that occur before or at the time the case would have been diagnosed in the absence of screening. (n.b., In a real population the time of diagnosis in the absence of screening for screen-detected cancers is not observable. However, it is available in the MISCAN life histories when screening is not offered.)

Note we presume for the cases, that screening ceases at diagnosis and thus, that the primary effect of the three exposure measures is to vary the latest time a control is eligible to be screened. Figure 3.1 presents the three exposure measures graphically when the case's cancer is diagnosed

Figure 3.1 Screening exposure criteria for controls when the case's cancer is clinically detected as a result of symptoms.

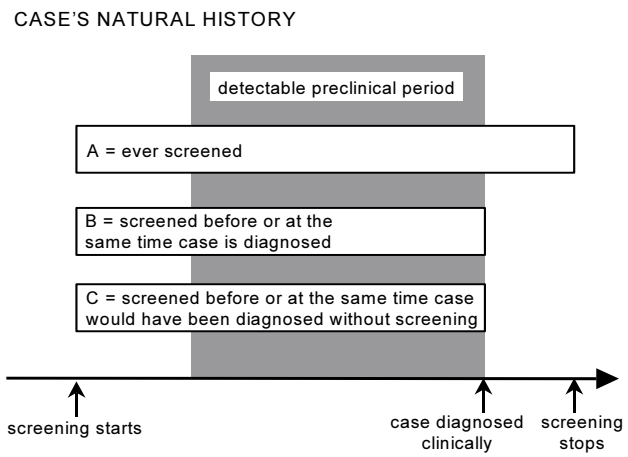
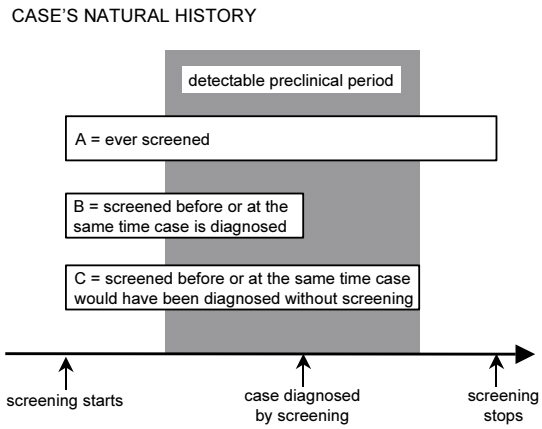


Figure 3.2 Screening exposure criteria for controls when the case's cancer is screen-detected.



clinically. Figure 3.2 presents the case when the case's cancer is screen-detected.

2. Type of Age Matching:
   One Year Age Window (i.e., a control's age must be within plus or minus one year of the case's age), or
   Birth Year (i.e., a control must be born in the same year as the case);
3. Time Scale:
   Calendar Time - the time of an event is the date of the event (e.g., the time of a woman's screen is the date on which the screen was done), or
   Chronologic Time - the time of an event is the age of the woman when the event occurs (e.g., the time of a woman's screen is the woman's age when the screen was done);
4. Eligibility Criteria for Age Matched Controls:
   a. none (i.e., all the age-matched controls are eligible)
   b. must be alive when the case died
   c. must be "cancer-free" when the case is diagnosed, (i.e., must not be diagnosed with breast cancer before the case is diagnosed), and
   d. must be both "cancer-free" when the case is diagnosed and alive when the case died.

There are 48 combinations of these factors for each MISCAN run. For both scenarios (the 100 runs with no benefit from screening and the 100 runs with benefit), the odds ratios obtained over the 100 runs for a particular combination are summarized using the weighted average of the estimates (with the weights being the inverse of the variance of the estimates). The standard error of the weighted average is calculated to assess precision.

Efficacy. For a MISCAN run the efficacy of being screened is given by

Efficacy = Mortality Rate(screened) / Mortality Rate(not screened)

where the Mortality Rate(screened) is the screened-for-cancer mortality rate among those who are screened at least once, i.e., those who accepted the screening offer at least once, and the Mortality Rate(not screened) is the screened-for-cancer mortality rate among these same individuals had they not been screened, i.e., their screened-for-cancer mortality rate in the absence of screening.

This information is available from the MISCAN life histories with and without screening being offered, and is used to calculate directly the efficacy obtained on each run. Note that this calculation is unnecessary for the scenario with no benefit from screening. Here the true efficacy is 1.00 by definition. For the scenario with a benefit from screening, the efficacy of being screened must be calculated for each of the 100 runs with a benefit. The mean of the efficacy of being screened for the 100 runs with a benefit from screening is used to estimate the true efficacy of screening when there is a benefit from screening.

results

The results are given for measures of exposure A, B and C by type of age matching (same birth year and one year age window), by time scale (calendar time and chronologic time), and by eligibility criteria (a, b, c, and d). Table 3.1 shows the summary odds ratio (and its standard error) for each combination of the factors for the 100 runs with no benefit from screening. Table 3.2 shows the summary odds ratio(and its standard error) for each combination of the factors for the 100 runs with a 50 percent reduction in the number of breast cancer deaths among those screen-detected. Note in tables 3.1 and 3.2 the standard errors of the estimated odds ratios indicate that the estimates are precise.

In the title of each table, the estimate of the true efficacy for that scenario is given. For table 3.1, the true efficacy for the 100 runs with no benefit is, as noted earlier, 1.00. For table 3.2, the estimate of the true efficacy of being screened for the 100 runs with a benefit from screening is 0.70.

Measures of exposure

For exposure A, which allows screening anytime, we observe that: 1) there is a substantial bias in favor of screening both when there is no benefit and when there is a benefit, and 2) the odds ratios are essentially the same for both types of age matching and for the two time scales, and only differ slightly for the eligibility criteria.

For exposure B an examination of tables 3.1 and 3.2 reveals a bias against screening. In table 3.1, all the odds ratios for the no benefit model are at best 21 percent greater than 1.00 and may be as much as 176 percent greater. In table 3.2, all the odds ratios for the benefit model where the estimated efficacy is 0.70 are at least 12 percent greater than 0.70 and may be as much as 86 percent greater. The combination of factors with the smallest bias is birth year age matching with the chronologic time scale and with one of the eligibility criteria b, c, or d. For the no benefit scenario the estimated odds ratio is 1.21 rather than 1.00 and for the benefit scenario the estimated odds ratio is 0.79 when the estimated efficacy is 0.70.

When exposure definition C is used, the bias against screening is greatly reduced. In particular, for the combination of birth year age matching with the chronologic time scale and with one of the eligibility criteria b, c, or d, the bias is nearly eliminated. In each of these circumstances, the odds ratio is close to the efficacy expected; with no benefit the estimated odds ratio is 1.01 versus the efficacy being 1.00, and with a benefit the estimated odds ratio is 0.70 when the estimated efficacy is 0.70.

Table 3.1 Estimated odds ratio (and standard error) given by measure of exposure by type of age matching by time scale and by eligibility criteria. No benefit from screening is assumed, true efficacy = 1.00.

Type of Age Match	Time Scale	Eligibility Criteria	Exposure Category A	B	C
Birth Year	Calendar	a	0.44(.008)	2.76(.055)	1.17(.020)
		b	0.41(.007)	2.75(.055)	1.15(.020)
		c	0.42(.007)	2.75(.055)	1.15(.020)
		d	0.41(.007)	2.76(.055)	1.15(.020)
	Chronologic	a	0.44(.008)	1.22(.021)	1.03(.017)
		b	0.41(.007)	1.21(.020)	1.01(.017)
		c	0.42(.007)	1.21(.020)	1.01(.017)
		d	0.41(.007)	1.21(.020)	1.01(.017)
Within One Year	Calendar	a	0.44(.008)	2.69(.053)	1.19(.021)
		b	0.41(.007)	2.67(.052)	1.17(.020)
		c	0.42(.007)	2.68(.052)	1.17(.021)
		d	0.41(.007)	2.68(.052)	1.17(.020)
	Chronologic	a	0.44(.008)	1.58(.028)	1.16(.020)
		b	0.41(.007)	1.56(.027)	1.14(.020)
		c	0.42(.007)	1.57(.027)1.	1.14(.020)
		d	0.41(.007)	57(.027)	1.14(.020)

Table 3.2 Estimated odds ratio (and standard error) given by measure of exposure by type of age matching by time scale and by eligibility criteria. A benefit from screening is assumed (a 50% reduction in the number of breast cancer deaths among the screen detected cancers), estimated true efficacy = 0.70.

Type of Age Match	Time Scale	Eligibility Criteria	Exposure Category A	B	C
Birth Year	Calendar	a	0.32(.006)	1.30(.026)	0.79(.014)
		b	0.29(.005)	1.28(.026)1.	0.77(.014)
		c	0.29(.005)	28(.026)	0.77(.014)
		d	0.29(.005)	1.28(.026)	0.77(.014)
	Chronologic	a	0.32(.006)	0.81(.0!4)	0.71(.013)
		b	0.29(.005)	0.79(.014)	0.70(.012)
		c	0.29(.005)	0.79(.014)	0.70(.012)
		d	0.29(.005)	0.79(.014)	0.70(.012)
Within One Year	Calendar	a	0.32(.006)	1.30(.026)	0.80(.015)
		b	0.29(.005)	1.28(.025)	0.78(.014)
		c	0.30(.005)	1.29(.025)	0.79(.014)
		d	0.29(.005)	1.29(.025)	0.78(.014)
	Chronologic	a	0.32(.006)	0.97(.018)	0.79(.014)
		b	0.29(.005)	0.96(.017)	0.77(.014)
		c	0.30(.005)	0.96(.017)	0.77(.014)
		d	0.29(.005)	0.96(.017)	0.77(.014)

age matching, time scale, and eligibility criteria
As noted above, these factors have little effect on the odds ratio when exposure A is used. However, for exposure measures B and C these factors do have an impact. For B the most obvious effect is the large bias against screening when the calendar time scale is used with either type of age matching. When chronologic time is used, the bias is reduced.
For exposure C the results are similar to those for measure B but with the amount of bias substantially reduced. In particular, when birth year age matching is chosen with chronologic time the bias is essentially eliminated.
The choice of eligibility criterion has only a small impact on the odds ratio. Criterion a yields the most distinct odds ratio. Its odds ratio stands apart from those for the three sets of results that require a specific vital status condition in addition to the age matching.
Thus, bias is essentially eliminated for the combination of exposure measure C with same birth year age matching with chronologic time and any one of the criteria b, c, or d. That is, if exposure C is used it appears that the case-control study can, in the absence of other biases such as self-selection bias, give an unbiased estimate of the efficacy of being screened.

discussion
measures of exposure
A substantial bias favoring screening when exposure A is used was not unexpected. Those without breast cancer are expected to have a higher rate of exposure to screening than those who develop breast cancer. This follows because those who do not develop breast cancer will continue to be invited for screening whereas those who develop breast cancer will not be invited for screening after their cancer is diagnosed. Hence measure A, which considers screening at any time as exposure to screening, will yield a greater exposure to screening for the women without breast cancer. Thus, the probability of being exposed to screening is greater for controls, most of whom will not have breast cancer, than it is for the "cases," all of whom were diagnosed with and died of breast cancer in the 15 year period. Exposure measure B avoids this bias by not considering as exposure to screening those examinations that take place after the case's breast cancer was diagnosed. This definition of exposure has been used in cancer screening case-control studies (Berrino et al. 1984; Friedman et al. 1995; Gill and Horwitz 1995; Hosek et al. 1996; Morrison 1982; Moss 1991; Sasco et al. 1986; Weiss 1983; Weiss and Lazovich 1996; Weiss et al. 1992). In our results, measure B removes measure A's bias in favor of screening but at the price of introducing a bias against screening. That this would occur had been suggested (Berrino 1993; Clarke and Anderson 1979; Selby et al. 1992). In particular, Berrino postulated that measure B would be biased against screening because a case is eligible to be screened until its cancer is detected either clinically or by screening, while controls matched to a case with a screen-detected cancer are only eligible to be screened until their matched case's cancer was screened detected. Thus the time interval in which controls matched to a case whose cancer was screen-detected is shorter than the time interval to the end of the case's preclinical period. To eliminate this bias, he suggested that the definition of exposure to screening be modified to include any screen up to and including the time the case's cancer would have been clinically diagnosed in the absence of screening. Figure 3.1 illustrates the situation when the case's cancer is clinically detected and figure 3.2 illustrates the situation when the case's cancer is screen-detected. In practice, because the end of the preclinical period is unknown, this measure cannot be used exactly. At best, it might be approximated using estimates of preclinical duration. However, in our microsimulation the detailed life histories for each individual both without and with screening are known. As a result, exposure could be evaluated when using this measure-measure C. The results support Berrino's hypothesis. Exposure measure B results in a bias against screening and measure C eliminates this bias. We plan to examine the use of approximations to exposure measure C in future research. At this point, our most important finding is that there is a bias against screening with exposure measure B that measure C eliminates.

age matching, time scale, and eligibility criteria
At first it is surprising that the type of age matching, a one year window versus same birth year, and the choice of time scale, calendar versus chronologic, can have a large effect on the odds ratio. For example, for age matching compare in table 3.1 the odds ratios for birth year age matching with those for within one year age matching when chronologic time and exposures B or C are used; and for time scale compare in table 3.1 the odds ratios for calendar time with those for chronologic time when birth year age matching and exposure B or C are used. However, on reflection it can be seen that at least some of the bias associated with these factors is due to the characteristics of the screening program that was simulated. Recall that the screening program started 1 January 1990, offered annual breast cancer screening to women from age 50 through age 70, and each year's screening examinations were scheduled to be on the women's birthdays which are uniformly distributed over the calendar year.
First consider the strong bias against screening that occurs when calendar time is used. This is seen for both types of age matching. Because in the

simulation birth dates are uniformly distributed over the calendar year and the birth days are the scheduled dates for the screening, it follows that controls are equally likely to be screened before or after their case's screening date. Thus, in any screening cycle, one-half of the case's controls will be screened after the case and thus, controls will tend to have a lower probability of being screened than the cases. In particular, in the cycle in which a case is detected, about one-half of the controls that are screened in that cycle would be considered as not exposed to screening if exposure measure B is used. If measure C is used, the reduction in the exposure to screening is smaller because the "cut-off" time for screening is not so closely tied to the date of detection. An extreme example of this calendar time bias is a situation in which measure B is being used and in which the case is detected at her first year of screening. Here about one-half of her controls who were screened in the first year would not be considered as exposed to screening (they are those whose screening date was later than the date the case was screened and detected with breast cancer) and thus, there would be a bias against screening. In contrast, the chronologic time scale, which uses the individual's age at the time of the event, removes this source of bias since all the controls of the same age would be considered exposed, even if their screens occurred after the date of diagnosis of the case.

As noted above, the randomness in birth dates and the use of the calendar time scale is of particular concern for exposure measure B where the date of the case's diagnosis ends the control's "ability to be screened" but is less important for exposure C since with C the cut off date for counting exposure for controls matched to screen detected cases is not fixed by the date that the case was screen detected.

Next consider the bias against screening that is seen using the one year window age matching with chronologic time. In the simulated screening program, when controls are age window matched there will be matched controls whose birth year is the year before the case's birth year. In such a situation, the controls first invitation to be screened will be either 1) for the screening cycle or round that follows the cycle in which the case was first invited or 2) for the same cycle as the case but at an older age as well as at a later date than the case's screening would be. In the former situation, cases will have one more screening round in which they can be screened than these controls will have. In the latter situation, if the case is screen detected, the control's screening in the cycle in which the case was detected would be after the case was detected and therefore, that screening would not be considered as exposure to screening under measure B(and it might not be considered as exposure under measure C). The end result is that if age window age matching is employed, cases will tend to have a higher probability of being screened than the controls. As an extreme example, consider a case born between 1920 and 1940 who had her breast cancer screen detected at the first screening cycle, in 1990. With one year window age matching there will be controls matched to the case who were born in the calendar year preceding the case's year of birth. Thus, for these controls the earliest they could have been screened is on their birthday in 1990. At this time the controls would be invited for screening at both a later date and an older age than the case. Thus, under measure B these controls will be considered as not exposed to screening (using measure C fewer of these controls would be considered as not exposed). Clearly, this results in a bias against screening. For a case detected at a later round of screening, it follows that the controls will also have a lower probability of being screened than the case but not as low as for the situation where the case is detected in the first round of screening. When same birth year age matching is used, this type of bias is greatly reduced. This follows since in the simulated screening program, yearly screening is offered and the screening cycle is a calendar year, from 1 January through 31 December each year. Hence, the cases and their same birth year controls are eligible for screening over the same screening rounds. Screening programs to be evaluated by a case-control study may experience poorer adherence to the scheduled periodic screenings and have different length screening cycles than the program simulated in this study, but similar large biases can occur if the chronology of the particular program is not well enough considered.

The choice of control eligibility criterion, a, b, c, or d, has little effect on the odds ratio. This is expected because in a normal risk population - simulated by MISCAN-there will be relatively few individuals eliminated by the necessity of being alive when the case died or cancer free when the case was diagnosed since few individuals experience these events. Although the differences are small, criterion a, which only age matches, yields the most distinct result with at most a very small bias against screening. The results are essentially the same for the last three criteria.

conclusion

These microsimulation results reveal a substantial bias against screening when the measure of exposure to screening used is definition B (i.e., one or more screens but only those that occur before or at the time the case's cancer is diagnosed). A number of studies have used this exposure measure and showed a substantial benefit from screening. This suggests either a strong benefit from the screening or a large self-selection bias in favor of screening, or both. It is difficult to determine which explanation holds or if

both do. We plan to investigate both the effect of the use of "more relevant periods of exposure", especially for the controls of cases with screen detected cancers, and of self-selection bias in future simulations. We also plan to explore the use of duration of time in the screening program as another matching variable.

The biases related to type of age matching and choice of time scale seem strongly related to the particular setting of screening within an organized screening program. These results may not strictly apply to settings other than the screening program scenarios which were simulated. This study did not consider a setting where the screening is "opportunistic," i.e., where screening is recommended but invitations to be screened are not offered, so that the time of screening is determined by the individuals themselves and is likely to be somewhat haphazard. In such a setting, similar biases can therefore be expected, but to be of a smaller magnitude.

Finally, although simulations are not truly reflective of reality and have strengths and weaknesses, these results make it advisable to consider the potential for bias from the measure of exposure, the type of age matching, and the time scale when undertaking a screening case-control study.

additional discussion

While the article of this chapter shows that there is an innate source of bias in the method of case-control studies on mortality effects of cancer screening, the question arises: can this bias be removed?

The problem would be solved if for screendetected cases the window in which exposure of the matched controls is measured would be extended with the lead-time of the case. This lead-time cannot be known for individual cases, but it can be estimated. A similar investigation as described in this chapter can be applied to an exposure measure that would extend exposure measure B with an estimate of the lead-time. This research is currently being carried out.

Another problem that remains is the bias caused by an association between attendance to screening and the risk of acquiring breast cancer and/or survival from breast cancer. It has been shown that this source of bias can be considerably larger than that described in the article of this chapter.(Gullberg et al. 1991) The approach of that article can also be applied to research bias in case control studies that arises from an assumed association between attendance to screening and the risk of acquiring breast cancer or survival from breast cancer. Also this research is currently being carried out.

4. A COMPARISON OF DISEASE SPECIFIC SURVIVAL OF PROSTATE CANCER FROM A POPULATION OF DEATHS AND FROM A POPULATION OF NEWLY DIAGNOSED CASES

abstract

purpose:
To investigate the validity of calculating prostate specific survival from a population of deaths occurring during a period of time.

materials and methods:
The stochastic simulation package Miscan was used to generate a large number of lifehistories of men with prostate cancer.

results:
The investigated way of calculating survival can lead to very different outcomes in comparison with the standard way of calculating survival. In an example with a Dutch population structure it leads to a strong underestimate of 25 years survival.

conclusions:
The investigated way of calculating survival is theoretically not valid and several probable changes in the population can produce results which are very different from the standard way of calculating survival.

introduction

The history of human prostate cancer if no intentionally curative treatment is applied, is unfortunately incompletely understood. A number of studies, of which (George 1988; Whitmore et al. 1991) are examples, have been published, all of which are biased by patient selection. The only prospective study reported (Johansson et al. 1992) is generally thought to underestimate the risk of prostate cancer progression and death. The average age in this study is 70 years, and poorly differentiated tumors were excluded for a period of time, which leads to the presence of these aggressive tumors of only 4% of the cases. It is not possible to quantify the bias caused by the selection effect, therefore a meta-analysis of the available literature data (Chodak et al. 1994) cannot correct for these problems, thus also leads to an underestimate of the risk of dying from locally confined prostate cancer.

In this particular situation the effort made by Aus and co-workers (Aus et al. 1995) to evaluate prostate cancer mortality from a population of all men who died in a certain geographic region during a given period of time seemed timely and promised to offer important data to the on-going discussion. However, the methodology used in the evaluation differs strikingly from the usual way of evaluating disease specific survival. Because of this the method has been criticized in this journal, but without explaining what is wrong with the method (Abrahamsson et al. 1996). Our study compares both approaches and considers the validity of the method of Aus by simulation of lifehistories with a known survival distribution and comparing the standard method for survival estimates with Aus' method. In this article the standard method will be called 'prospective' and the method used by Aus 'retrospective'.

material and methods

The populations in this study consist of lifehistories simulated by a Miscan model for evaluation of prostate cancer screening. Miscan is a stochastic simulation package for simulating individual life histories of which the principles have been described elsewhere (Habbema et al. 1985). So far there have only been published results from application of the package in evaluation of cervical cancer screening and breast cancer screening (de Koning et al. 1995a; de Koning et al. 1991; van Ballegooijen et al. 1992a; van Oortmarssen et al. 1992). The model used for the present paper is developed for the evaluation of prostate cancer screening. As the model is intended primarily for evaluation of the ongoing randomized screening trial in Rotterdam (as part of the European Randomised study of Screening for Prostate Cancer), some characteristics of the model apply to the Dutch situation. Characteristics of this study are:

The model reproduces age specific incidence of prostate cancer in the Netherlands in the period of 1989-1991 (NeCaRe 1991) as well as mortality by age of Dutch males from other causes than prostate cancer. (CBS 1996). Both prostate cancer incidence and mortality from other causes are assumed to be constant over time in absence of screening.

It simulates a full dynamic population, which means that births and deaths occur as in a normal population, contrary to many other simulation models which reproduce a birth cohort in which age of the population is completely parallel to calendar time. There are two variants in the demographic pattern of the model: One which reproduces the present Dutch age distribution of males, the other simulates a population of which births are uniformly distributed over time.

For reasons of simplicity there has been chosen an arbitrary disease specific survival from prostate cancer in a situation without screening: men diagnosed with prostate cancer have an exponential distribution of cause specific lethality from prostate cancer (that is 1 - survival) with a mean duration of 20 years. This leads to a survival curve which is more or less similar to the survival curve presented in the paper of Aus and co-workers for all M0 cases. Also for simplicity reasons survival does not depend on age.

All survival rates presented in the study are prostate cancer specific survivals. They are either calculated by the standard method from the usual population of newly diagnosed cases (prospective) or from the population which consists of those men dying between 1988 and 1990 (retrospective) who have a diagnosis of prostate cancer. Survival rates were calculated from the simulated lifehistories by the SPSS statistical program according to the Kaplan-Meier method (Kaplan and Meier 1958).

In a variant of the model survival rates strongly improve for cases detected in the year 1983 or later. This is modeled by starting screening in the year 1983 which is applied to the population of men who are 55 to 69 years old with a screening interval (and duration of a screening round) of 4 years. The screening in this model variant is assumed to only lead to earlier detection of prostate cancer (lead time), but not to influence the moment of death in comparison with a situation when no screening would have occurred. This means that screening increases incidence at the screeningsages, but does not change mortality. The assumptions on preclinical duration and sensitivity of the screening test advance the diagnosis by an average time of 4 years per cancer detected at screening (lead time), consequently the survival of screen-detected cases is increased by an average of 4 years.

The simulated population of present Dutch age structure consists of 12,995 life histories for prospective survival; from this population 1578 died in the period 1988-1990 therefore are in the population for retrospective survival and 1573 for retrospective survival on cases diagnosed from 1958 (the start of the Swedish Cancer Registry). The simulated population with uniformly distributed births without screening consists of 22,641 life histories for prospective survival, of which 2802 for retrospective survival. The simulated population with uniformly distributed births with screening introduced in 1983 consists of 15,031 life histories for prospective survival up to 1983, and 13,846 between 1983 and 1990, and 3235 for retrospective survival.

results

Figure 4.1 shows the results from prospective and retrospective survival analysis with Dutch population

Figure 4.1 Prospective and retrospective survival in a population with Dutch age structure and the influence of excluding cases diagnosed before 1958 on retrospective survival

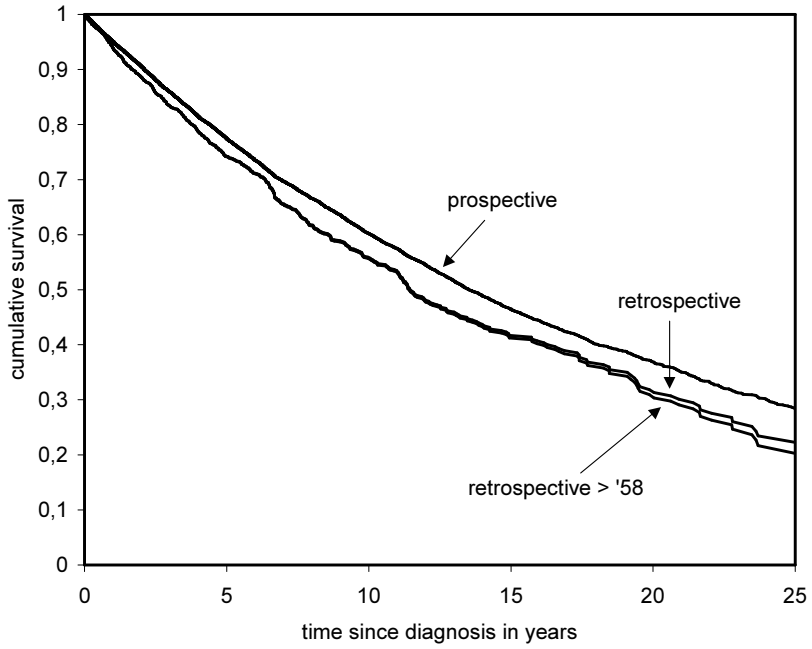
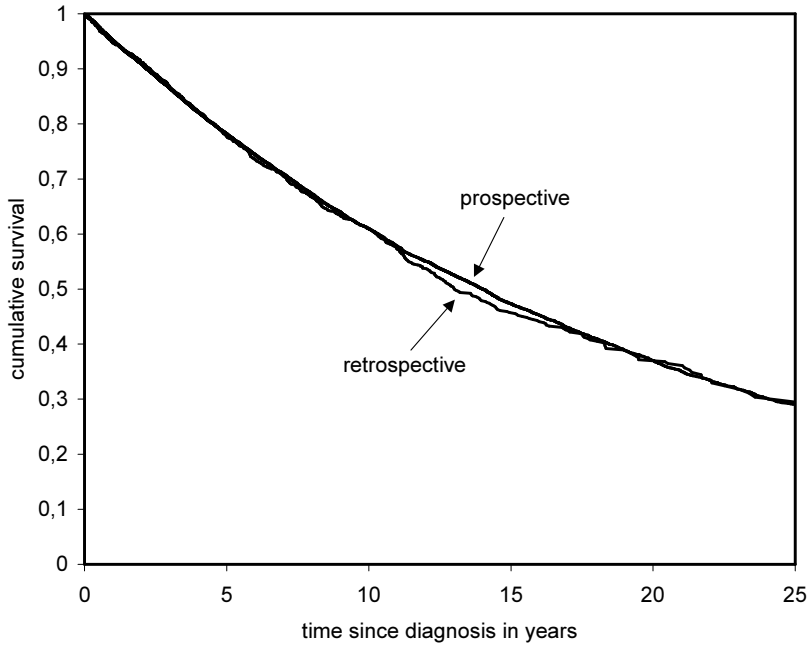
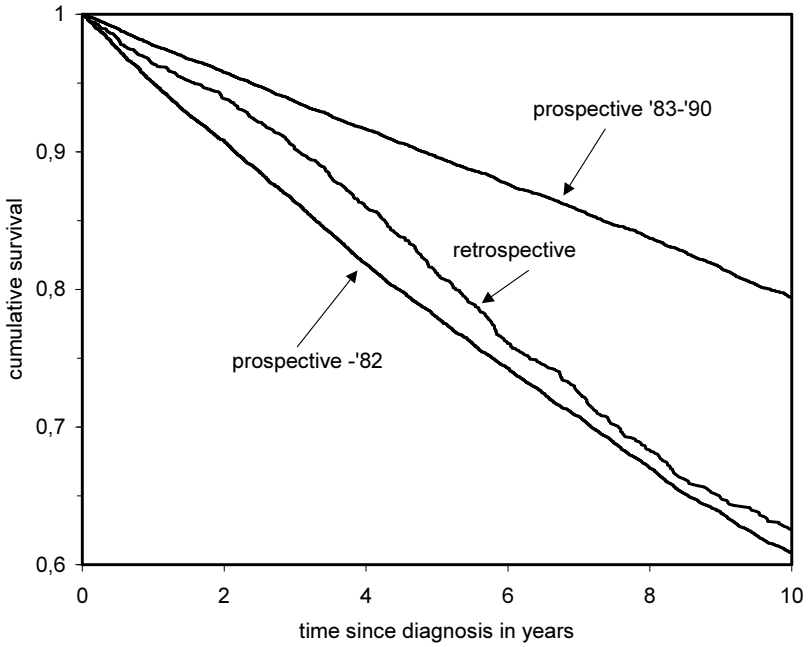


Figure 4.2 Prospective and retrospective survival in a population in which births are uniformly distributed over calender time



structure. There is clearly a considerable difference between the two methods. The figure also shows the difference between retrospective survival when applied to all men who die between 1988 and 1990 with a diagnosis of prostate cancer and when applied to the same population but excluding those with a diagnosis of prostate cancer of before the start of the Swedish cancer registry in 1958. This exclusion only affects very few cases (5 out of 1578 in this particular simulation) therefore there is hardly any difference in calculated survival as long as the population under follow-up is relatively large. Only after about 20 years of follow-up, when the remaining population under follow-up is less than 4% of total population, there is a discernable difference because of this exclusion. And then still this difference is small in comparison with the difference between prospective and retrospective survival.

Figure 4.3 The influence on prospective and retrospective survival of a sudden improvement of survival due to introduction of screening in 1983 without effect on mortality



As should be expected, the results of the prospective survival analysis reproduce the model assumptions: 29% for 25 years survival (expected: $e^{-25/20}$), in comparison with 23% for retrospective survival analysis.

When making the same comparison in a population with evenly distributed births, the difference between the two methods disappears as is shown in figure 4.2. The small difference between the two methods is to be attributed to random error in the simulated outcomes. The 25 years survival is 29% in both cases.

When (assumed not effective) screening is introduced in 1983 to the model with evenly distributed births, 5 years prospective survival increases from 78% for the period up to 1982 to 90% for the period 1983-1990. Retrospective 5 years survival with cases dying in the same period is 81%, therefore reproducing not even half of the increase in prospective 5 years survival. (fig. 3)

discussion

The population of prospective survival consists of a cohort of people who enrol due to diagnosis of the studied disease. All members of this population are at risk for dying from the disease during all of their time of follow-up, therefore they are constituting a 'population at risk'. In retrospective survival membership of the study population is not only determined by the diagnosis of the disease, but also by the time of death, that means that the terminal event of the study determines enrolment into the study. This is in contradiction with a proper study into a cause-effect relationship of appearance of disease and dying from that disease in which the effect to be studied should not influence the membership of the study.

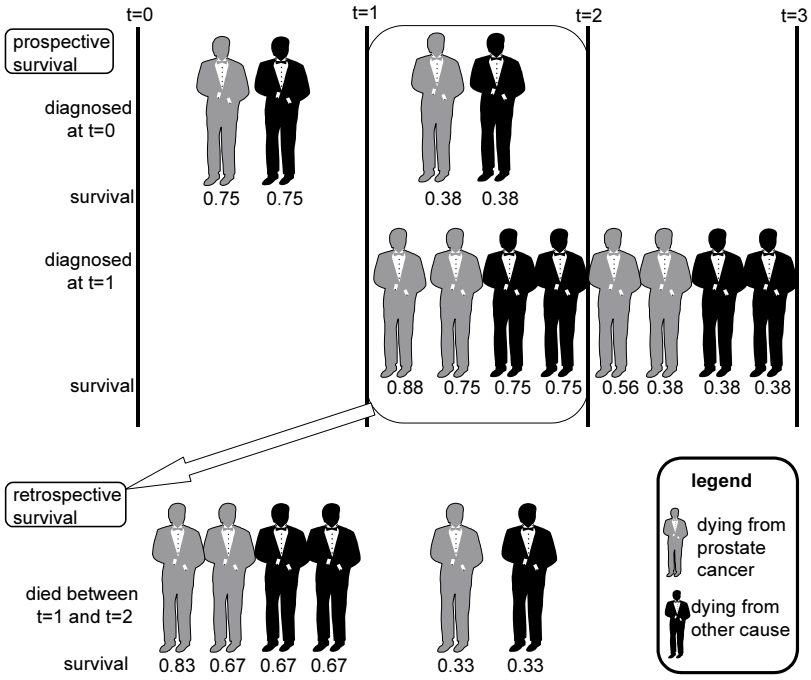
In a complete steady state when all parameters concerning numbers of births over calender time, mortality from other causes than the disease, incidence of the disease and survival from the disease remain constant over time, then retrospective survival gives the same answers as prospective survival. As soon as one or more of these parameters change over time, the retrospective survival analysis becomes unreliable which may lead to either an over- or underestimate of real survival.

The difference between prospective and retrospective survival in a Dutch population is to be explained as follows: Prospective survival is calculated by decreasing the proportion surviving with the mortality ratio for each time that someone dies from prostate cancer from start of follow up since diagnosis. The mortality ratio is the number of deaths divided by the number of people at risk.

Retrospective survival is comparable to that, except that the denominator is not the population at risk of dying from the disease, but consists of those people who have been diagnosed before the time of follow up under consideration. In a population such as the Dutch, where the number of elderly people is increasing, the number of new cases of prostate cancer as well as the total number of deaths is

increasing over time. The denominators used in retrospective survival analysis are smaller than the real population at risk, because they consist of the population of (elderly people) before the time under consideration.
Underestimation of the real denominators leads to overestimation of mortality and thus to underestimation of survival.
This mechanism can be illustrated in figure 4.4 by a simplified example: Consider the population where at time = 0 year there are 4 new cases of prostate cancer. During the first year of follow up first one man dies from prostate cancer, then one dies of other causes. During the second year the same events occur in the same order, after which all have died. The disease specific prospective survival after the first event is 3/4, after the second event it is 3/4 * 1/2 = 0.375.

Figure 4.4 A simple example to explain the difference between prospective and retrospective survival in a population in which the number of newly diagnosed cases of prostate cancer increases over time while survival is constant



Assuming that the population doubles in the next year and for the rest all remains the same then: At time = 1 year, there are 8 new cases of prostate cancer, during the first year, firstly two men die from prostate cancer, then two die of other causes, again the same during the second year. The survival of these 8 cases diagnosed at time = 1 year is exactly the same as from the 4 cases diagnosed at time = 0 year.
Now we take the population of deaths occurring from time = 1 year to time = 2 year, which is according to the retrospective method of survival analysis. Those are the deaths during the second year of follow-up of the cases occurring at time = 0 year and the ones during the first year of follow-up of the new cases diagnosed at time = 1 year. Calculating disease specific survival from this population gives: During the first year since diagnosis 2 men die when the population in follow-up is 6, therefore survival is 4/6, during the second year 1 man dies when the population in follow-up is 2, therefore cumulative survival is 4/6 * 1/2 = 0.333, which is lower than that of prospective survival. When assuming a population with uniformly distributed births as well as constant all cause mortality, the population pyramid exactly follows the overall survival from birth. When incidence of the disease under consideration is also constant, then the population in the retrospective survival study parallels that of the prospective survival study, leading to equality of the outcomes of the prospective and retrospective method.

The retrospective method does not pick up much of a recent increase in survival due to early detection because most of the denominator (also during the first years of survival) is from old cases.
If an increase of survival in a stable population would not be due to early detection but to improved treatment, then in the first period of follow up after diagnosis prospective and retrospective survival will be very similar. But also in that case in later years of follow up the denominators will be lower in retrospective survival because it consists of men who have been diagnosed in earlier years than the year from which current mortality is considered. Due to the increasing survival, more men have already died from prostate cancer before the moment of counting the denominator than in the true population at risk, therefore the population at risk is underestimated, leading to a lower survival outcome.
Comparing retrospective and prospective survival in a case of changing survival over time is not quite straightforward because they do not use the same distribution over time of diagnosed cases. In a population with increasing survival which is otherwise stable, retrospective survival will not become lower than the oldest prospective survival, therefore in that case it might be argued if retrospective survival is wrong.
Grönberg (Grönberg 1995) has attributed the observed difference between retrospective and prospective survival in northern Sweden only to an increase in survival. That may be true, but it is not

certain as long as no other possibilities of bias as described in this article have been ruled out.
Using the population structure by age of Göteborg between 1988-1990 might have led to another difference between prospective and retrospective survival.
Other changes in the population, not taken into account in this article, can lead to further differences between prospective and retrospective survival. An example of that is a change in age specific mortality from other causes than prostate cancer, which is likely to have occurred in the Gothenburg population from 1958 to 1990. A decreasing mortality in a population with evenly distributed births would lead to an overestimate of survival due to a sampling bias very similar to what is described earlier in this article.
Because survival is intended to represent what is to be expected from the time of diagnosis, and not what has happened before dying, the retrospective method should not be considered as a valid estimate for survival.

5. AMBIGUITIES IN CALCULATING CANCER PATIENT SURVIVAL: THE SEER EXPERIENCE FOR COLORECTAL AND PROSTATE CANCER

abstract

background

When estimating survival from cancer registry data, there are several decisions to be made concerning record selection and method of calculation. For instance, should one use relative or disease-specific survival, how to handle 'autopsy only' cases and cases with unknown cause of death. This study evaluates the effects of such decisions for survival from colorectal or prostate cancer in the SEER program.

methods

The research population consists of cases of colorectal cancer and prostate cancer as reported to a SEER cancer registry from 1973 through 1994. Several alternative methods of estimating survival were evaluated. One method of disease-specific survival is chosen as index scenario, and variants in the selection of cancer cases and the method of estimating survival are compared to this index scenario.

results

In general, the differences between the different survival estimates are small for colorectal cancer and somewhat larger for prostate cancer. The most substantial differences concern disease-specific versus relative survival for prostate cancer and the inclusion or exclusion of individuals with more than one cancer for both cancers investigated.

conclusions

There is no single best method for calculating cancer survival. Practical suggestions for calculating cancer survival when using SEER program data are derived from this study.

introduction

When estimating survival from cancer registry data, there are several decisions to be made concerning record selection and method of calculation. This paper only concerns estimating net survival that is intended to show the influence of having the disease in question. (Estève et al. 1994) The possibilities of estimating net survival depend on specific information available from the registry in question. The SEER program data contain a sufficient level of detail to evaluate how a variety of decisions lead to different survival estimates: which records of primary cancers should be considered, should one use relative or disease-specific survival, and in case of the latter, what is the right definition of disease-specific death, how to handle 'autopsy only' cases, 'death certificate only' cases and cases with unknown cause of death or with no known death certificate? The best choice may depend on the purpose of the survival estimate. Survival estimates can, for instance, be used for studying time trends, in order to monitor the eventual effect of dissemination of a new therapy or for studying the effect of stage at diagnosis or that of socio-economic status. Our research was initiated in response to the problem of estimating the possible effects of screening, but we will also discuss alternative purposes that may lead to selecting a different method of estimating survival. Specifically, this study focuses on 5 year survival from colorectal cancer and 10 year survival from prostate cancer. Colorectal cancer was chosen as a common cancer in which relatively few difficulties concerning survival estimates are to be expected. Prostate cancer is also common but, in contrast to colorectal cancer, can be expected to be accompanied by more ambiguities concerning survival calculation because the cause of death of a prostate cancer patient is often less clear and mortality from the cancer does not decrease as rapidly with time since diagnosis as in most other cancers.

material and methods

The research population consists of cases of colorectal cancer and prostate cancer as reported to a SEER cancer registry from 1973 through 1994. (National Cancer Institute 1997) A case of colorectal cancer is defined by the codes for 'primary site' of C18.0 through C20.9 and prostate cancer is defined by code C61.9 (Percy et al. 1990a).

The index scenario was chosen so that the other methods of calculating survival used in this study could be derived by simple alterations. This index scenario is disease-specific survival with the following exclusions (in parentheses is the exact coding of the exclusion): carcinoma in situ (historic stage = 0), cases with unknown survival (survival time = 9999), cases with more than one primary cancer at the time of last follow-up (sequence number > 0), cases without a death certificate or listing available to SEER (underlying cause of death is 7777) cases with a death certificate but without a coded cause of death (underlying cause of death is 7797), cancers detected at autopsy only (type of reporting source = 6), cases known by death certificate only (type of reporting source = 7), and individuals coded to be of 'other race' according to Race recode B (race = 8 through 98) or of unknown race (race = 99), leaving individuals identified as white, black, American Indian, Alaskan Native, Chinese, Japanese, Filipino and Hawaiian.

Table 5.1 shows the number of primary cancers present in the SEER program, the numbers of

primary cancers excluded for the index scenario of survival calculation and the number of cases remaining. Some cancers are excluded for more than one reason, therefore the remaining number of cases plus the sum of the numbers of exclusions is larger than the initial number of cancers.
Disease-specific survival is calculated by the actuarial method with one month intervals since survival time is coded in months by SEER. All causes of death due to cancer (underlying cause of death 140.0 through 209.9) are counted as death due to the disease under study.
Variants in the selection of cancer cases and the method of estimating survival are described by what is changed relative to this index scenario.

relative survival
The index scenario uses disease specific survival and treats deaths from causes other than the disease in question as 'lost to follow up' as the case does not contribute any further to the person years at risk of dying from the disease. Relative survival is calculated by dividing observed survival by expected survival.(Ederer et al. 1961) Expected survival, which would have applied if the patient would not have the disease in question, is derived from the total population of people from which the diagnosed cases were drawn.
Here, relative survival was calculated by using SEER*Stat, a PC based survival system provided by NCI (National Cancer Institute 1997) with the same exclusions as the index scenario.

number of cancers in one individual
In the index scenario all individuals with a history of more than one cancer at any site were excluded from survival calculations. Two variants were studied: the first considered each primary colorectal cancer or prostate cancer, and the second considered only the first invasive colorectal or prostate cancer in an individual. An individual is defined as a set of records with the same participant number and the same case number. If more than one cancer of the same site was diagnosed at the same time, then only the one with the worst stage is considered where the order of increasing severity of stage is chosen to be: localized, regional, unstaged, distant.

disease-specific death
In the index scenario, the widest definition for disease-specific death is used: all cancer deaths are considered to be disease-specific death. As long as only individuals are considered with no more than one primary cancer at the end of follow-up, this seems reasonable. If someone is attributed to have died from a different cancer than the only one that was diagnosed, it can be regarded as probably an error on the death certificate. In order to check the extent to which this influences the estimates, two variants of disease-specific death are considered: a narrow definition in which disease-specific death from colorectal cancer is defined as cause of death from colon cancer or rectal cancer (153.0 through 154.1) or from unspecified organs of the tractus digestivus (159.0 through 159.9); death from prostate cancer then is defined as simply from prostate cancer (185.0 through 185.9). The wider definition of disease-specific death for colorectal cancer as well as prostate cancer includes death specified as metastases (195.0 through 199.9).
Because of possible interaction between the three definitions of disease-specific death (narrow, wider and index scenario/widest) and three selections according to number of primary cancers at the end o follow-up (one primary only, only the first primary of the cancer site and all cancers of the site in question), all nine combinations are considered.

table 5.1 Number of primary cancers in the SEER program from 1973 through 1994, numbers of exclusions for the index scenario of survival calculation and remaining number of cases

	colorectal cancer	prostate cancer
total number of cases	265,463	225,358
exclusion because of:		
carcinoma in situ	14,188	374
unknown survival	3,419	4,605
sequence number > 0	67,240	38,167
no death certificate	6,463	5,111
unknown cause of death	499	561
autopsy only and cancer death	523	817
autopsy only and death from other causes	833	2,441
death certificate only	2,062	1,347
other race	1,006	689
unknown race	1,577	3,601
remaining cases in index scenario	179,545	174,861

cases detected at autopsy
Some primary cancers are registered as detected only at autopsy. Most of these cancers are not considered to have caused the death of the individual. They should best be regarded as prevalent cases of disease that apparently did not yet cause any harm to the individual. However, some of these cancers are attributed to have caused the death of the individual. In one variant the latter category of cancers is included in the survival estimate and attributed a survival of 0 months.

cases known by death certificate only
Death certificate only cases are cases in which the patient has a cause of death of cancer, but follow

back with the hospital and physician did not yield evidence of a cancer diagnosis. We included 'death certificate only' only if cancer is actually mentioned as the underlying cause of death. Since the date of diagnosis is not known, the survival time is unknown. In order to check what is the maximum possible error by excluding these cases (as in the index scenario), in two variants these cases are included by assuming to have disease-specific survival of 0 months on the one hand and on the other hand a survival for at least the survival period in consideration (5 or 10 years). The latter poses a problem for attributing the right age of diagnosis. For reasons of simplicity, the age of death is chosen as the age of diagnosis.

cases without death certificate or without coded cause of death
In the index scenario, cases that did not have a death certificate or where there was no coded cause of death are not included in the survival estimate. In two variants, these cases are included in the estimate: one assumed all deaths were disease-specific and one assumed all deaths were from other causes.

individuals of 'other race' or of unknown race
As a variant to the index scenario, cancer cases from individuals of 'other race' or of unknown race are included in the cancer survival estimate.

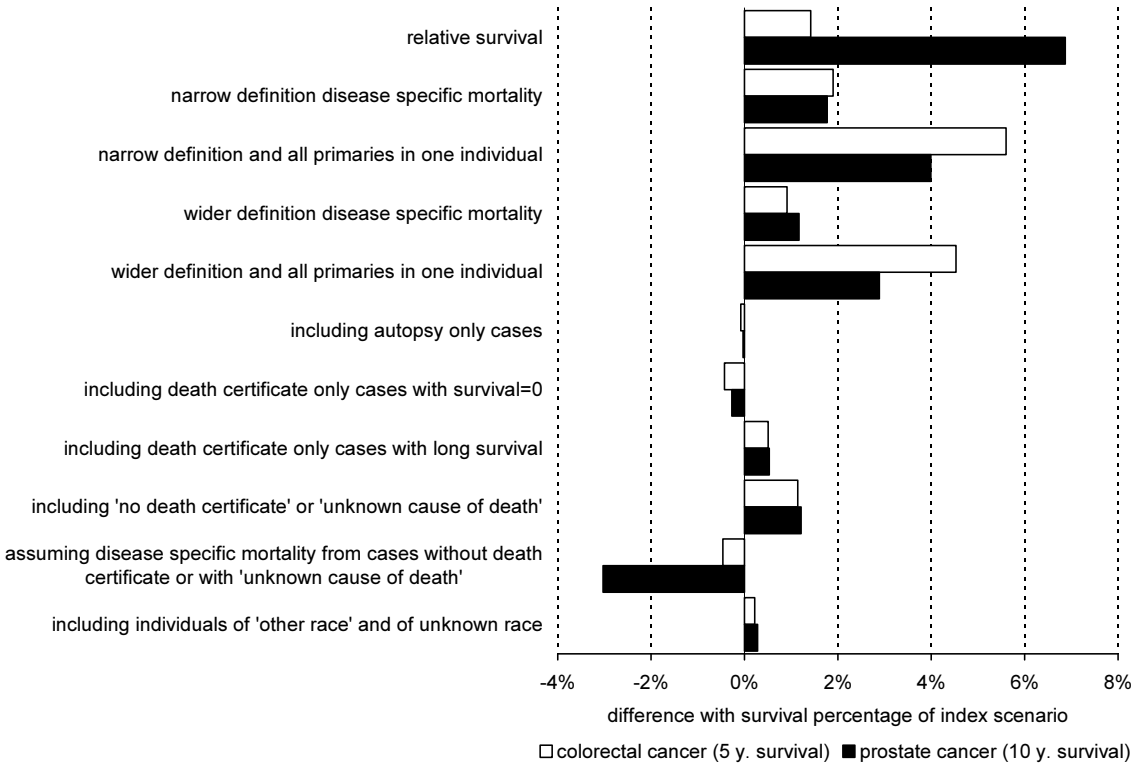
results
An overview of the outcomes of the different methods of survival estimation are given in table 5.2 for colorectal cancer and in table 5.3 for prostate cancer.
In general, the differences between the outcomes for colorectal cancer are small. The differences are somewhat larger among the unstaged cases and among older individuals. The differences between the outcomes for prostate cancer are larger than for colorectal cancer. There they do not concentrate so much in the unstaged cases but also for prostate cancer the differences are larger among older men than among younger ones.

disease-specific versus relative survival
The differences between disease-specific and relative survival of colorectal cancer are small. Under age 75 they are equal except for the fact that, when compared with disease-specific survival, the staged cases have a slightly higher relative survival compensated by a lower relative survival of unstaged cases. Among older individuals relative survival tends to be higher than disease-specific survival.
For prostate cancers the differences are much larger. In all stages and age categories for which survival was estimated, relative survival shows higher outcomes than disease-specific survival, except for cases with distant metastases under age 75. The differences are larger than 10% for some categories.

number of cancers in one individual
Among the 25,386 people who had a first colorectal cancer followed by diagnosis of one or more further cancers (sequence number = 1), survival is 69.1% in comparison to 51.6% for those who had only one primary cancer at the end of follow-up (sequence number = 0). The 34,397 cases of colorectal cancer that are diagnosed among individuals with a previous history of cancer (sequence number > 1) show a slightly lower 5 year survival of 49.1%.
If only first colorectal cancers are taken into account these figures do not change much: 69.5% for people who had a first colorectal cancer followed by diagnosis of one or more further cancers and 47.6% for people who already have a history of another type of cancer.
The 17,184 cases of prostate cancer among people who had a first prostate cancer followed by diagnosis of one or more further cancers have a 10 year survival of 39.5% (compared to 59.4% for the people only one cancer), the 18,710 cases with previous history of cancer have a survival of 41.7%.
Table 5.2 and 5.3 show an expected tendency for survivals from cancers with more favorable stages to be more influenced by other cancers than that of more unfavorable stages.

disease-specific death
Of the individuals with colorectal cancer selected according to the index scenario who die within 5 years from cancer, 94.5% have a cause of death according to the narrow definition of disease-specific death (colon or rectal cancer, or tractus digestivus unspecified), 3.0% are attributed to have died from metastases and 2.5% from other cancer death. Of colorectal cancer cases where the cancer is the first among more than one cancer and who die within 5 years from cancer, only 64.1% die from colon or rectal cancer, or tractus digestivus unspecified. For cases with a previous cancer, this is 75.4%. The fraction attributed to dying from metastases does not depend strongly on sequence number: 3.5% for first among more cases as well as for cases with a previous cancer. Of course, the fraction dying from other cancers is much higher among those with more than one cancer: 32.4% for first cases and 21.1% for later cases.
Of the individuals with prostate cancer selected according to the index scenario who die within 10 years from cancer, the figures are very similar to those of colorectal cancer: 94.4% are classified as having died from prostate cancer, 2.1% from metastases and 3.5% from other cancer death.
Because prostate cancer survival in this study is estimated for a follow up period twice as long as that

Figure 5.1 Differences between survival percentage of variants with the index scenario of 5 year colorectal cancer survival and 10 year prostate cancer survival



for colorectal cancer, there is more opportunity to die from other cancers for those individuals who have more than one cancer. Of prostate cancer cases where prostate cancer is the first among more than one cancer and who die of cancer within 10 years, only 23.6% die from prostate cancer and for prostate cancer cases with a history of other cancer: 41.2%. The fraction attributed to dying from metastases among individuals with prostate cancer is clearly higher if one has more than one cancer: 5.8% for first cases and 4.7% for later cases. Also here, the fraction dying from other cancers is much higher among those with more than one cancer: 70.6% for first cases and 54.1% for later cases.
Because few individuals who die with only one cancer are attributed to dying from metastasis or dying from another cancer than they actually have, survival estimates do not become a lot higher when applying the narrow definition of disease-specific death as compared to the index scenario of considering all cancers deaths to be disease-specific: 53.5% versus 51.6% for 5 year survival from colorectal cancer and 61.2% versus 59.4% for 10 year survival from prostate cancer. For individuals who have more than one cancer, narrowing the definition of disease-specific death has a much greater influence. Five year survival from colorectal cancer is 79.3% versus 69.1% for first cases and 58.8% versus 49.1% for later cases. For 10 year survival from prostate cancer the effect is even larger: 80.1% versus 39.5% for first cases and 66.6% versus 41.7% for later cases, so much that also for prostate cancer cases survival appears to be better among those with more than one cancer.
Tables 5.2 and 5.3 also show that selection of only first colorectal or prostate cancers makes little difference.

cases detected at autopsy and cases known by death certificate only
Including cancer cases detected at autopsy, whose deaths are attributed to the given cancer, lowers survival estimates. However, the number of cases involved is relatively small, so the effect on survival estimates is not substantial as shown in tables 5.2 and 3. Also the number of 'death certificate only' cases is relatively low, thus their inclusion is of little influence on survival estimates. Of course all 'death certificate only' cases are of unknown stage, therefore there is no possible influence on survival of known stages.

cases without death certificate or without coded cause of death
The number of individuals with cancer who died but for whom no death certificate is available or the death certificate doesn't give a cause of death is a lot larger than 'autopsy only' and 'death certificate only' cases, therefore the possible influence on survival estimates is larger. Including them by attributing

table 5.2: 5 year survival from colorectal cancer in the SEER program from 1973 through 1994 according to different methods of estimation, differences of the odds and the odds of the index scenario of more than 20% are in italics

	all ages	<75 years of age					75+ years of age				
	all stages	all	localised	regional	distant	unstaged	all	localised	regional	distant	unstaged
index scenario	51.6%	54.3%	86.6%	55.8%	6.0%	43.8%	46.5%	77.9%	48.9%	4.4%	22.9%
relative survival	53.0%	54.3%	87.6%	56.2%	6.1%	39.9%	49.5%	*84.1%*	52.9%	4.9%	*19.2%*
number of primaries per individual											
all primaries	53.2%	56.3%	84.1%	57.1%	6.8%	44.9%	47.7%	74.0%	49.3%	5.0%	24.3%
first invasive colorectal cancers	53.3%	56.3%	85.0%	57.3%	6.8%	45.3%	47.7%	75.1%	49.7%	5.0%	24.2%
definition of cancer death											
narrow definition*	53.5%	56.0%	87.6%	57.2%	6.7%	46.2%	48.8%	79.4%	51.1%	*5.3%*	24.7%
all primaries	*57.2%*	*59.8%*	87.6%	60.1%	*8.0%*	50.2%	*52.6%*	79.5%	*53.8%*	*6.3%*	*28.5%*
first invasive colorectal cancers	*57.2%*	*59.8%*	88.5%	*60.3%*	*7.9%*	*50.8%*	*52.5%*	80.5%	*54.2%*	*6.2%*	*28.4%*
wider definition*	52.5%	55.1%	87.2%	56.4%	6.3%	45.0%	47.7%	78.8%	50.1%	4.8%	23.9%
all primaries	56.1%	58.8%	87.0%	59.3%	7.4%	48.8%	51.3%	78.7%	52.8%	5.7%	27.4%
first invasive colorectal cancers	56.1%	58.8%	88.0%	59.5%	7.3%	49.3%	51.3%	79.8%	53.1%	5.7%	27.4%
case selection											
include autopsy only cases	51.5%	54.2%	86.6%	55.7%	5.9%	43.7%	46.4%	77.7%	48.8%	4.4%	22.9%
include DCO* with 0 survival	51.1%	54.1%				40.4%	45.8%				20.1%
include DCO* with > 5 years survival	52.1%	54.5%				*49.3%*	47.7%				*37.0%*
include missing/unknown COD*											
as not dead of disease	52.7%	55.3%	87.0%	56.8%	6.6%	45.8%	47.9%	78.7%	50.4%	5.1%	24.6%
as dead of disease	51.1%	53.9%	85.9%	55.3%	5.9%	43.5%	46.0%	76.6%	48.3%	4.6%	22.8%
include 'other race'/unknown race	51.8%	54.6%	86.8%	55.9%	6.0%	45.5%	46.6%	77.9%	49.0%	4.5%	23.3%

DCO = death certificate only; COD = cause of death; narrow definition = COD 153.0-154.1, 159.0-159.9; wider definition = narrow + COD 195.0-199.9

table 5.3: 10 year survival from prostate cancer in the SEER program from 1973 through 1994 according to different methods of estimation, differences of the odds and the odds of the index scenario of more than 20% are in italics

	all ages	<75 years of age					75+ years of age				
	all stages	all	localised	regional	distant	unstaged	all	localised	regional	distant	unstaged
index scenario	59.4%	63.7%	79.0%	59.7%	13.9%	57.6%	50.2%	63.7%	42.4%	13.3%	46.9%
relative survival	*66.3%*	*70.3%*	*87.7%*	*69.5%*	13.5%	*62.6%*	*57.3%*	*71.3%*	*52.6%*	14.1%	*58.2%*
number of primaries per individual											
all primaries	55.2%	59.8%	*72.4%*	57.1%	13.9%	53.7%	45.8%	*56.9%*	38.9%	12.8%	42.1%
first invasive prostate cancers	55.2%	59.8%	*72.4%*	57.1%	13.9%	53.7%	45.8%	*56.9%*	38.9%	12.9%	42.1%
definition of cancer death											
narrow definition*	61.2%	65.2%	80.4%	60.9%	14.9%	58.7%	52.6%	66.3%	44.3%	14.5%	48.9%
all primaries	63.4%	67.4%	81.5%	63.2%	*16.3%*	61.4%	*54.9%*	*67.9%*	*47.1%*	15.6%	50.5%
first invasive prostate cancers	63.4%	67.4%	81.5%	63.2%	*16.4%*	61.4%	*54.9%*	*67.9%*	*47.2%*	15.6%	50.5%
wider definition*	60.6%	64.7%	79.9%	60.6%	14.5%	58.2%	51.8%	65.5%	43.7%	14.0%	48.4%
all primaries	62.3%	66.4%	80.4%	62.4%	15.8%	60.5%	53.7%	66.6%	46.1%	15.0%	49.5%
first invasive prostate cancers	62.3%	66.4%	80.4%	62.4%	15.8%	60.5%	53.7%	66.6%	46.1%	15.0%	49.5%
case selection											
include autopsy only cases	59.4%	63.7%	78.9%	59.7%	13.9%	57.6%	50.2%	63.7%	42.3%	13.3%	46.9%
include DCO* with 0 survival	59.2%	63.6%				56.7%	49.7%				44.0%
include DCO* with > 5 years survival	60.0%	63.8%				60.0%	52.1%				*59.7%*
include missing/unknown COD*											
as not dead of disease	60.6%	64.7%	79.7%	60.9%	15.3%	58.9%	51.9%	65.2%	44.1%	14.8%	48.5%
as dead of disease	56.4%	61.2%	76.1%	57.5%	13.4%	54.6%	46.1%	58.5%	39.5%	12.3%	42.1%
include 'other race'/unknown race	59.7%	64.0%	79.1%	59.9%	14.1%	58.5%	50.6%	63.9%	42.5%	13.4%	48.2%

DCO = death certificate only; COD = cause of death; narrow definition = COD 185.0-185.9; wider definition = narrow + COD 195.0-199.9

their death as non disease-specific leads to higher survival estimates and including them as disease-specific death leads to lower estimates than the index scenario.

individuals of 'other race' or of unknown race
Individuals who are coded to be of race 'other' have a markedly higher 5 or 10 year survival from respectively colorectal cancer (61.6%, n=786) and prostate cancer (65.7%, n=594) than individuals with a specified racial code.
Individuals of unknown race have a very high survival: 89.9% for 5 year survival from colorectal cancer (n=1033) and 94.5% for 10 year survival from prostate cancer (n=3387). Therefore including those two groups in the survival estimate leads to a slight increase as shown in tables 5.2 and 5.3. At the end of known follow up in SEER, 87.0% of colorectal cancer cases of individuals of unknown race are considered as being still alive versus 36.3% of individuals with specified race, for prostate cancer this is 95.8% versus 53.9%. Among individuals of unknown race, 52.5% of colorectal cancer is diagnosed in stage 'localized' (versus 33.3% among individuals with specified race), and 25.2% is unstaged. For prostate cancer, 41.5% is staged localized among individuals of unknown race (versus 58.1% among individuals with specified race) and 48.3% of cases from unknown race is unstaged.
Figure 5.1 gives an overview of the differences between survival percentage of variants with the index scenario of 5 year colorectal cancer survival and 10 year prostate cancer survival.

discussion
In 5 year survival from colorectal cancer each of the different methods of estimating survival which were used in this study gave quite similar results. Differences in 10 year survival from prostate cancer are more substantial and perhaps best explained by the fact that the underlying cause of death is often unclear and the period of follow up since diagnosis was longer.
Net survival is an adjustment of observed survival. In all variants of survival estimation methods considered here, the underlying observed survival is not varied, so differences are due to the adjustment to net survival for mortality from other causes. That's why one can expect differences between the methods to be larger with increasing mortality from other causes. This is the case for prostate cancer compared to colorectal cancer where follow up time is longer and cases are older.

disease-specific versus relative survival
Relative survival does not require accurate registration of causes of death. But it is crucial for relative survival that mortality in the background population is representative for the risk to die among cases from other causes than those to be attributed to the disease.
An obvious advantage of disease-specific survival is that it only needs information on diagnosed cases. But for disease-specific survival it is crucial that the registered cause of death from the specific disease is indeed counting all deaths, but no more than those, which are to be attributed to the disease.
The SEER program usually uses relative survival in order to avoid problems in death misclassification and because of ambiguities in definitions (Brown et al. 1993; Percy et al. 1990b).
The differences between disease-specific and relative survival, particularly in prostate cancer, may be due to bias in either of the two methods. On the one hand, too many deaths of patients with prostate cancer may be attributed to prostate cancer as the underlying cause of death, leading to an underestimate of net survival. On the other hand, expected survival (used for relative survival) may be underestimated due to the use of life tables based on the entire U.S.A. instead of the areas of the SEER program and/or because individuals diagnosed with prostate cancer may have a lower risk of dying from causes unrelated to prostate cancer. An underestimate of expected survival implies a too large correction for mortality from other causes, thus an overestimate of net survival. These biases in expected survival seem unlikely to cause all of the 10% difference in survival estimates since that would imply around 20% lower mortality from other causes than average for individuals under age 75 (expected 10 year survival is 61%). There are signs that the misclassification of the cause of death is occurring on a substantial scale. The introduction of PSA screening has led to a huge, temporary, increase in incidence of prostate cancer due to the early detection of prevalent cases which was accompanied by an increase in registered prostate cancer mortality. (Feuer et al. 1999) If there is indeed a tendency to attribute death to prostate cancer for someone with a history of that disease but who is not actually dying from it, this would explain both at least part of the differences between relative and observed survival as well as the coincidence in increasing incidence and mortality of prostate cancer around the time of introduction of PSA screening.

definition of disease-specific death
It appeared from this study that the problem of inaccurate specification of cancer death is rather limited, leading to only a 2% difference in survival, depending on the definition of disease-specific death. The question remains which definition is closest to the truth. There are 1839 cases of colorectal cancer selected for the index scenario who are attributed to have died from cancer but not from a cause in the narrow definition of disease-specific death nor from

metastases, and 1287 of such prostate cancer cases. Among those 3126 cases, the largest number of deaths is attributed to lung cancer (942 cases with code 162). One may wonder if this is due to an incorrect specification of cause of death, or to primary lung cancer that has not reached the SEER program as reportable incidence. The next two largest categories among colorectal cancer cases concern cancer in the liver and the pancreas, and among prostate cancer, cancer in the bladder and the pancreas. These seem likely candidates for misspecification. But the next larger categories concern stomach cancer, ovarian and prostate cancer among colorectal cancer cases and colorectal cancer among prostate cancer cases, which again do not seem likely to be misspecified.
One might consider a thorough review of all of the mortality codes and consider whether they are likely candidates for misspecification. In that case, one might also consider some causes of death which are not due to cancer but may be due to therapy, such as sepsis.

number of cancers in one individual
When estimating net survival for individuals with more than one cancer, not all cancer deaths should be regarded as disease-specific, but only a more narrow definition of disease-specific death is useful.
From a perspective of therapy, contrary to what is observed, having more than one cancer would be expected to lead to worse survival because co-morbidity may be a reason for less rigorous treatment.
The observed differences can be explained by a bias due to case selection: individuals who die quickly after the diagnosis of their first cancer have little opportunity of having a next cancer and cancer cases which are the first of a sequence are selected for having longer survival. There is an indicator for such effect from the stage distributions, 33.3% of colorectal cancer cases among people with only one cancer are localized, 47.6% of cases that are first in a sequence and 43.4% of cases with a previous cancer. For prostate cancer a similar effect is found: 58.1% of cases among people with only one cancer are localized, 65.8% of cases that are first in a sequence and 62.5% of cases with a previous cancer. But this result may also be (partly) due to the fact that detection of more than one cancer in an individual is stimulated by a higher cancer awareness as well as by follow up visits after treatment of the first cancer, both leading to earlier detection of another cancer which might have remained undiagnosed otherwise, at least within the period of observation.
Whatever the causes of the observed differences, it does not appear to be correct to discard all cancer cases among individuals with more than one cancer if one is interested in an accurate estimate of net survival. For instance, if one is interested in estimating prognosis at diagnosis, there is no way to know in advance if the individual will be diagnosed with another cancer later on, therefore in that case one should at least also consider individuals who have the cancer of interest as first in a series of more than one diagnosed cancer.

cases detected at autopsy
Though the number of colorectal and prostate cancer cases detected at autopsy are low, one has to make a decision on how to deal with them. From the perspective of determining prognosis for someone diagnosed with cancer, clearly these cases should not be included in the survival estimate because for cases detected at autopsy, the question of prognosis is not meaningful. But from a perspective of evaluation of early detection, they should be included because earlier diagnosis and treatment can improve their prognosis.

cases known by death certificate only
The probability of observing a case identified by death certificate only increases if one or more of the regular reporting sources to the registry are not involved, in particular if the diagnosis does not involve a laboratory for histology, and/or a hospital admission of the patient. It can be expected that cases of this type have a different survival than other types of cases and it is probably worse. One can estimate the theoretical maximum overestimate that is made due to disregarding 'death certificate only' cases by attributing them a survival of 0. This study shows that for all cases of colorectal cancer and prostate cancer in SEER, the maximum of the overestimate due to leaving out 'death certificate only' cases is at most quite small. At older ages and when selecting unstaged cases, the possible bias is larger.
Trying to estimate the theoretical maximum underestimate by assuming that 'death certificate only' cases have a long survival is problematic. It is not known to which age category the individual should be attributed. Also, one should then consider the possibility of an unknown quantity of unregistered cases of cancer which are still alive. However, leaving these out decreases the survival estimate, thus giving a conservative maximum underestimate.

cases without death certificate or without coded cause of death
Some members of the SEER population are registered as 'no death certificate' or 'unknown cause of death' by the time they die. Leaving out cases coded as such from the survival estimate, leads to an overestimate of net survival because the members of that part of the population who are still alive are not left out. As a variant one can include them and

attribute them to disease-specific death. Of colorectal cancer cases selected according to the index scenario who die within 5 years with a known cause of death, 76.7% die from cancer. If one assumes that this is the same for people of whom no death certificate is available, then the survival estimate that assumes that cases with no death certificate are dying from disease-specific cause is closer to the truth than when assuming they are dying from other causes. In prostate cancer survival it is not so clear because only 48.8% of prostate cancers cases who die within 10 years die from prostate cancer. An important reason for the absence of a death certificate is probably that the individual has moved out of the region of the SEER program. It is conceivable that moving is associated somehow with the probability of dying from colorectal or prostate cancer.
For those who have an unknown cause of death the percentage who have in fact died from the cancer in question may be quite different, but that group is considerably smaller than the group without death certificate.

individuals of 'other race' or of unknown race
The main reason for the high survival rates of individuals of unknown race is probably that the death certificate often yields an individual's race, consequently many individuals who have died are selected out of this category. This mechanism also causes a slight underestimate of survival in all other racial categories.

observed survival
This study has not considered possibilities of error in the observed (or crude) survival. The number of cases in the register with unknown survival is limited and therefore their maximum influence on estimated survival is small. Besides that, cases that are lost to follow up are not likely to have a radically different survival from other cases after the moment of loss. Also, the number of cancer cases that are diagnosed but do not reach the registry appears to be limited considering how few cases are known to the registry by death certificate only.

conclusion
There is no single best method for calculating survival from cancer in the SEER program. Different methods can give different outcomes, but for most variants considered the differences are small. The most substantial differences in this study concern disease-specific versus relative survival for prostate cancer and the inclusion or exclusion of individuals with more than one cancer for both cancers investigated.

suggestions for calculating cancer survival in the SEER program

1. Since disease-specific and relative survival are subject to different forms of bias of variable magnitude, it is advisable to apply both methods, to compare the outcomes and to try to explain the differences. The preferred method may be different depending on cancer site and the study objectives.

2. Various definitions of disease-specific death are possible. When considering individuals with only one cancer, the definition is not crucial. It is advisable to consider which codes of cancer death should be included as disease-specific death and which should not, depending on their probability to be inaccurately coded deaths with respect to the cancer for which survival is estimated. Also some codes for non cancer death may be considered such as sepsis, particularly if therapy for the cancer involves a substantial operation risk.

3. One should seriously consider including individuals with more than one cancer. If one estimates net survival by means of disease-specific survival according to the narrow definition of disease-specific death, this is not problematic. If one wants to include all cancer deaths as disease-specific, one solution may be to estimate the difference between using the narrow and wider definition from the individuals with only one cancer and extrapolate that difference to those with more than one cancer. In relative survival this is not a problem if one assumes that incidence at different cancer sites is not correlated.

4. Cases detected at autopsy that are not attributed to disease-specific death are generally not relevant for survival estimates. If one is evaluating early detection, the other 'autopsy only' cases should be included as having survival 0, but if one needs the estimate for prognosis at time of diagnosis, one should not include them. However the influence of 'autopsy only cases' is likely to be small.

5. Cases with unknown survival time, cases known by death certificate only, cases known to have died but without death certificate available and cases without coded cause of death from the death certificate should be excluded from a baseline survival estimate. The percentage of these cases should be noted.

6. The maximum bias due to exclusion of cases known by death certificate only can be shown by assuming that these cases have a 0 survival. The maximum bias due to exclusion of cases without death certificate or without coded cause of death can be determined by assuming that these cases die from disease-specific death (or simply include them in relative survival). But judging from the account of their percentages, it is likely to be not worthwhile bothering about these biases.

7. If one is not categorizing survival by race, then survival from cases among individuals of 'other race' or of unknown race should be included in the estimate.

6. COST EFFECTIVENESS OF SHORTENING SCREENING INTERVAL OR EXTENDING AGE RANGE OF NHS BREAST SCREENING PROGRAMME: COMPUTER SIMULATION STUDY

abstract

objective

To compare the cost-effectiveness of two possible modifications to the current U.K. screening programme; shortening the screening interval from 3 years to 2 years and extending the age of invitation to a final screen from 64 to 69.

design

A computer simulation model which first simulates life histories for women in the absence of a screening programme for breast cancer, and then assesses how these life histories would change as a consequence of introducing different screening policies. The model was informed by screening and cost data from the NHS breast screening programme in the North West region of England.

results

Compared with the current breast screening programme both of the alternatives are predicted to lead to an increase in the number of deaths prevented and in the number of life years saved. The model calculates that the current screening policy costs £2,522 per life year gained, extending the age range of the programme £2,612 and shortening the interval £2,709 per life year gained. The marginal cost per life year gained of extending the age range of the screening programme is £2,990 and of shortening the screening interval is £3,545.

conclusions

If the budget for the NHS breast screening programme were to allow for two more invitations per woman, the MISCAN model predicts substantial mortality reductions would follow from extending the age range screened or reducing the screening interval and suggests that the difference between the two policies is so small that, depending on the outcome measure considered, either can be preferred.

introduction

In 1988, the NHS breast screening programme on the recommendation of an expert committee chaired by Sir Patrick Forrest, began screening women aged 50 - 64 years every three years. However, the committee also concluded that the optimum frequency of screening and the age range likely to benefit from breast screening was still undetermined (Forrest 1987). This study uses a computer simulation package MIcrosimulation SCreening ANalysis (MISCAN) to compare the cost-effectiveness of two possible modifications to the current UK screening programme; shortening the screening interval from 3 years to 2 years and extending the age of invitation to a final screen from 64 to 69.

description of the MISCAN model

A full description of the MISCAN model has already been published (van Oortmarssen et al. 1990b). In brief, the model first simulates life histories for women in the absence of a screening programme for breast cancer, and then assesses how these life histories would change as a consequence of introducing different screening policies.

The natural history of breast cancer is modelled as a progression from no breast cancer through pre-clinical cancer to clinical disease. Women reside in the first state (no breast cancer) before entering one of five pre-clinical states. There is an in situ state, and four invasive states according to the tumour size (T-status); (5mm (T1a), >5-10mm (T1b), >10-20mm (T1c), and >20mm (T2+). A cancer may be detected at a screening, or become clinically apparent in any one of these states, or if undiagnosed progress to the next pre-clinical state. The two end-states of the model are death from breast cancer and death from other causes.

In the initial parameterisation of the MISCAN model, data from the Dutch screening trials at Utrecht and Nijmegen were used to provide estimates of the mean duration of the pre-clinical phase for women in different age groups and the mean duration of cancer in each of the five preclinical states. The dwelling time of a cancer in each pre-clinical state is assumed to follow an exponential distribution and the rate with which cancers progress from the pre-clinical to the clinical state is inferred from the observed incidence and stage distribution of clinically diagnosed cancers in the population under study.

When modelling the performance of a screening programme, key parameters include the mean duration of the screen-detectable phase, the sensitivity of the test and the improvement in prognosis for screen-detected cancers. The mean duration of pre-clinical screen-detectable period assumed in the model was based on data from the Dutch screening projects at Nijmegen and Utrecht and varied from approximately 1.8 years at age 35 to 6.2 years at age 70. The sensitivity of the screening test is assumed in the model to be the probability of detecting a cancer in the pre-clinical screen detectable state and for women aged over 50 is fixed as: 0.4, 0.65, 0.8, 0.9 and 0.95 for in situ disease, T1a, T1b, T1c and T2 + respectively. The improvement in prognosis for screen-detected cancers was derived from the results of the Swedish breast screening trials (de Koning et al. 1995a).

applying the MISCAN model to a UK population

The North West region of the UK has a population of 4.1 million and is covered by 5 NHS breast screening programmes. The largest of these, the Manchester Breast Screening Programme has now screened over 120,000 women and reported cancer detection rates similar to those elsewhere in the U.K. (Chamberlain et al. 1993). The number and size of cancers detected at a first and second screen and the occurrence and size of interval cancers in this programme have been used to inform the model. Estimates of screening and diagnostic costs are based on this programme assuming that two view mammography is used at a first screen and single view mammography at subsequent screens. Treatment costs are derived from a variety of sources, but primarily the Christie Hospital NHS Trust in Manchester. Full details of the costing, including sensitivity analysis, are provided in Street et al 1996) (Street et al. 1996). Both costs and effects are discounted at 6%.

Table 6.1: Detection rates per 1,000 screened

	first screenings		second screenings*	
age	observed	modelled	observed	modelled
50-54	5.5	4.6	-	3.7
55-59	5.6	5.6	-	4.0
60-64	7.1	7.6	-	5.5
all ages	6.0	5.9	4.7	4.6

* Age specific rates were not available

Table 6.2: Stage distribution by T status %

	first screening		second screening	
stage	observed	modelled	observed	modelled
DCIS	14.2	14.9	17.0	15.1
T1a	8.6	9.1	3.8	13.0
T1b	31.5	31.0	29.6	38.8
T1c	31.0	30.2	33.9	25.6
T2+	14.7	14.8	15.7	7.5

In order to simulate the life histories of women with breast cancer before a screening programme is introduced, the model requires information on the age, stage distribution and survival of women with breast cancer. Neither the pre-screening stage distribution nor stage specific survival rates prior to the introduction of screening were available for the North West's population. However, the pre-screening stage distribution in Scotland (Scottish Cancer Therapy Network 1996) and in East Anglia (personal communication J. McCann, East Anglian Cancer Registry) was similar to that of the control population in the Utrecht screening trial and it was therefore assumed that the pre-screening stage distribution in the North West was similar to that was used in the initial parameterisation of the MISCAN model. The stage distribution in women aged 50-69 at diagnosis in the Utrecht control population was: 4.6% DCIS, 1.5% T1a, 6.3% T1b, 32.6% T1c and 55% T2+. Having assumed this stage distribution, stage and age specific survival rates were derived by fitting the North West's observed 1987 mortality rate to the North West's observed 1987 incidence rate. This produced an overall 5 year survival for women aged 50-59 and women aged 60-69 of 67% and 68% respectively. A life table describing the probability of dying from causes other than breast cancer in the North West's population was used to derive the number of life years gained per breast cancer death prevented.

The model as originally parameterised was unable to simulate the detection rate and stage distribution observed at first screening in the North West. More small cancers were observed in the North West than were predicted by the MISCAN model. This discrepancy was resolved by assuming a longer screen detectable pre-clinical phase for small tumours in the North West. When it was assumed that small tumours (less than 10mm) dwelt in a screen detectable phase for twice as long as that used in the initial Dutch model the model adequately fitted the detection rate and stage distribution observed at first screening in the North West.

This model was used to simulate for the population of the North West the effects and costs of three screening programmes; first the current UK screening policy, in which women aged between 50 and 64 are invited for screening every three years; second, extending the age range of women screened from 64 to 69 years, but continuing to invite women every three years and finally reducing the screening interval from 3 to 2 years while maintaining the current age range. Attendance for screening was assumed to decline 0.5% for each year of age from 74.2% at age 50 to an attendance of 67.9% at age 70; attendance at repeat invitations is assumed to be 78 % higher among those who attended the previous invitation. Each screening programme was assumed to run for 27 years.

results

The final model adequately predicted the rates of screen detected cancers observed at the first and second screening round, the interval cancer rates observed after a first screen and the stage distribution observed at the first screen but not the stage distribution observed at the second screen (Tables 6.1-6.3).

Table 6.3: Interval cancer rates after first screening round per 10,000 screened

months after screen	observed	modelled
0 - 11	5.5	5.4
12 - 23	9.2	9.8
24 - 35	14.9	13.0
total	29.5	28.2

A summary of the costs and effects of the three screening policies, compared with a baseline of no screening, is provided in Table 6.4. This suggests that the current North West screening programme reduces mortality by 12.8%, preventing 4,079 deaths during the 27 year period; this is equivalent to 66,187 life years gained or 12,251 life years discounted to present values.

Extending the programme to age 69 reduces mortality by 16.4%, preventing 5,311 deaths during the 27 year period, equivalent to 78,221 life years gained or 15,161 life years discounted to present values. Reducing the interval to two years reduces mortality by 15.3% preventing 4,880 deaths, equivalent to 81,322 life years gained or 14,987 life years discounted to the present values.

Table 6.4: Overview of the effects and costs of three screening policies (costs in millions of £)

	present programme	extension to age 70	two year interval
Effectiveness (no discounting)			
mortality reduction *	12.8%	16.4%	15.3%
deaths prevented/year *	147	188	175
deaths prevented (total)	4079	5311	4880
life years gained	66187	78221	81322
Costs (6% discounted)			
screening	26.8	34.7	36.9
diagnosis (screening)	17.4	21.1	19.7
diagnosis (outside screening)	-11.1	-14.1	-12.9
primary therapy	6.2	9.1	7.1
adjuvant therapy	-1.5	-2.1	-1.8
follow up	2.9	4.0	3.5
advanced disease	-9.8	-13.1	-11.8
total	30.9	39.6	40.6
Effectiveness (6% discounted)			
deaths prevented	1229	1636	1457
life years gained	12251	15161	14987
Cost-effectiveness (6% discounted)			
costs per death prevented (£)	25142	24205	27865
costs per life year gained (£)	2522	2611	2709
marginal cost-effectiveness per life year gained (£)		2990	3545

* in a steady state

The cost of the current programme £30.9 million increases to £39.6 million if the age range of the programme is extended and to £40.6 million if the screening interval is reduced. The majority of these resources are spent on screening and investigation of women recalled with a suspicious screening film but some resources are saved because of the reduced diagnostic and treatment costs in women who would otherwise have presented symptomatically.

These data suggest that the cost of a life year gained by screening when costs and benefits are discounted at 6% (derived by dividing discounted life years gained by the cost of the programme) is £2,522 in the current programme, £2,611 if the age range of the programme is extended and £2,709 if the screening interval is shortened. The impact of changing the current screening policy is best summarised by comparing the marginal cost effectiveness of the two alternative policies which is calculated by dividing the difference in total costs of the current and proposed policies by the difference in life years gained. The marginal cost per life year saved of extending the age range of the screening programme is £2,990 and of shortening the screening interval £3,545.

The cost effectiveness analysis was also conducted using, respectively, the lower and upper unit cost estimates for screening, diagnosis and treatment and alternative discount rates. Under all scenarios considered the current programme implies a lower marginal cost per life year saved or death prevented than the two policy options. (A more detailed exposition of the cost and sensitivity analyses are available in Street et al 1996 (Street et al. 1996).) Two alternative

models using different assumptions about the length of the pre-clinical detectable phase and the size distribution of tumours at presentation were also explored and the relative outcomes on cost-effectiveness were the same. These models were rejected, however, because they did not fit all the available data as adequately as the model described.

discussion

The MISCAN model was developed and refined over a number of years at the Erasmus University Rotterdam. It has been validated using data from the Netherlands (van Oortmarssen et al. 1990b) and Sweden (de Koning et al. 1995a), assumptions underpinning the model have been evaluated by others (Brown and Fintor 1993; Chamberlain et al. 1993) and the results from the model have been used to evaluate screening programmes in several European countries (Beemsterboer et al. 1994; de Koning et al. 1991; van Ineveld et al. 1993). In order for the model to simulate the detection rates and stage distribution observed at first screening in the North West's population, it was necessary to assume for smaller tumours, a longer pre-clinical detectable phase than was estimated from the Dutch pilot projects and the Dutch national screening programme. A longer pre-clinical detectable phase is in accordance with a lower threshold of detection of breast cancer at screening.

The model adequately simulated the number of cancers occurring in the interval between screens and those detected at a second screen but it predicted a better stage distribution for the latter than was observed. This tendency for the MISCAN model to predict a better stage distribution at repeat screens than is observed has been reported before (Paci et al. 1995) and is being investigated by the Erasmus team. It is unlikely, however, that this discrepancy substantially affected the conclusion, given that a better stage distribution at repeat screens is modelled in all policy options and that the overall reduction in mortality predicted for each screening policy option is not greater than those reported from the randomised trials of breast screening.

In cost effectiveness analysis of programmes whose main effect is to extend life the usual measure of benefit is life years gained. Compared with the current breast screening programme both of the alternatives evaluated offer an improvement in the effectiveness of the programme; both are predicted to lead to an increase in the number of life years gained and in the number of deaths prevented. Establishing a preference between the two proposed policies depends on making choices about which outcome measure is chosen, whether discounting is undertaken, and whether costs are considered. If no discounting of health effects is undertaken and costs are ignored, extending the age range appears preferable when measuring outcomes by deaths prevented but in terms of life years gained reducing the screening interval is the more effective policy. Whether either of the proposed changes to the programme is cost-effective depends on the value the NHS is willing to place on incremental improvements in the effectiveness of the programme. On the basis of a comparison of the current programme with a hypothetical baseline of no screening, (assuming a discount rate of 6% applied to all costs and outcomes), the implicit incremental costs embodied in the current programme are £25,142 per death prevented, and £2,522 per life year saved. If these values represent an upper limit on the amount which the NHS is willing to pay, the cost effectiveness of the proposed changes depends on the outcome measure. The incremental cost per life year saved is higher with both proposed changes than the current programme but extending the age range for screening, offers an incremental increase in deaths prevented at a cost which is lower than that implicit in the current programme. If this is an appropriate measure of outcome, then extending the age range is unambiguously cost effective. Looking specifically at the relative cost-effectiveness of the two policy changes, the incremental cost per additional death prevented and per life year saved is slightly lower with extending the age range.

In conclusion, if the budget for the NHS breast screening programme would allow for two more invitations per woman, the MISCAN model predicts that the difference between extending the age range screened or reducing the screening interval from three to two years is so small that, depending on the outcome measure considered either can be preferred. This conclusion remained consistent when using upper and lower estimates of the costs of screening, diagnosis, treatment, and when varying the discount rate of costs and benefits.

7. UNAIDED VISUAL EXAMINATION OF THE CERVIX VERSUS PAP-SMEARS FOR EARLY DETECTION OF CERVICAL CANCER

abstract

Unaided visual inspection of the cervix can be applied as an alternative to pap smear testing for the prevention of mortality from cervical cancer. We investigated the comparative cost-effectiveness of these two modalities for early detection of cervical cancer for a situation in India.

We estimate that per 100 pap smears taken at age 45 the expected number of life years gained by prevention of mortality from cervical cancer is 7.3. Screening an equal number of women two times by unaided visual examination at ages 39 and 51 is expected to gain slightly more life years. The estimated cost of a pap smear in India is 37.3 rupees. The cost of a visual examination is 11.8 rupees, but due to low specificity this examination generates much more diagnostic costs.

Therefore our conclusion is that pap smear taking may be more cost-effective than unaided visual examination of the cervix, also in a developing country like India.

introduction

In many developing countries, cancer of the cervix is the most important cancer with respect to the risk of acquiring the cancer as well as the risk of dying from it. Both these risks can be strongly reduced by an effort to detect curable precursors of invasive cervical cancer applying cytological screening by pap smears. This form of secondary prevention is widely applied in developed countries, where the risk for cervical cancer is usually not very high, while it is applied relatively rarely in developing countries where this risk is much higher. An important reason for the absence of cervical cancer screening in developing countries is that taking pap smears is costly and requires a fairly high degree of organisation which also involves cytological laboratories.

One particular alternative to cytological screening, unaided visual examination, is seriously being discussed and investigated (Bhargava et al. 1993; Nene et al. 1996; Sehgal et al. 1991; Singh et al. 1992; Sujathan et al. 1995; Wesley et al. 1997).This screening test is a direct visual inspection of the cervix in order to look for macroscopic abnormalities which possibly signify (precursors of) cervical cancer. It does not require a cytological laboratory, the procedure as such is less costly than cytological screening and the outcome of the test is immediately available. The main disadvantages of visual examination are that its sensitivity for finding precursors of cervical cancer as well as its specificity are considerably lower than for the pap smear test. This paper gives a quantification of the advantages and disadvantages of visual examination and pap test in order to evaluate which screening method is more cost-effective. The quantification of the test characteristics of visual examination is based on published data.

material and methods

The MISCAN model for underlying natural history of cervical cancer and the impact of screening, which has been used for this analysis, simulates individual life histories consisting of demographic, epidemiological and screening aspects.

Demography in the model concerns births by calendar year and mortality from other causes than cervical cancer. These data are derived from the age specific death rates in India 1987 (Registrar General of India 1995) and from the age distribution of the female population of Bangalore as estimated for 1985 (Parkin et al. 1992).

The disease model distinguishes the following consecutive preclinical disease states which are detectable by screening: dysplasia, carcinoma in situ (CIS), micro-invasive cancer (IA) and macro-invasive cancer (IB+). The dwelling time in each of these disease states follows a Weibull distribution with a coefficient of variation of 0.55. The mean duration in dysplasia is 4 years (standard deviation 2.2 years), in CIS 7.8 years (s.d. 4.3), in IA 2 years (s.d. 1.1) and in IB+ until clinical diagnosis 3 years (s.d. 1.6). These durations are based on the assumption that the natural history of cervical cancer (precursors) is similar in Bangalore and in the Netherlands, except for the state IB+ which is assumed to be 3 years in Bangalore in comparison to 1.9 years in the Netherlands because clinical diagnosis is assumed to occur later in Bangalore. (van Ballegooijen et al. 1992a; van Ballegooijen et al. 1995; van Oortmarssen et al. 1992) Incidence and survival of women with cervical cancer by age follows that of the situation in Bangalore over the years 1981-1989 (Nandakumar et al. 1995). The age standardised incidence rate based on the World Standard Population (ASR-W) is 28.8 (Nandakumar et al. 1995), which is much higher as for instance in the European Union where the ASR-W for 1990 was 10.2 (Black et al. 1997) or the United States with an ASR-W of 7.8 per 100,000 life years during 1990-1994 (National Cancer Institute 1997). The onset of dysplasia is chosen so that the target clinical incidence is reproduced.

The sensitivity of the pap smear test is assumed to be 80% for dysplasia and CIS and 85% for invasive cancer. Specificity is assumed to be 98%. (van Ballegooijen 1998a)

The test characteristics for screening by visual examination in the baseline model are adjusted according to the experience in Maternal and child health centres in Delhi (Singh et al. 1992). The sensitivity of visual examination is assumed to be 40% for dysplasia and CIS and 75% for invasive cancer, and specificity 90%. When detected with (a precursor of) cancer by screening in stage IA or earlier everyone is assumed to be cured. When detected by screening in stage IB+ the probability to be cured decreases linearly over time from 100% if one is detected just after entry in stage IB+ to 0% if screen detection is just before the time of diagnosis in a situation without screening. Here cure is defined as treatment leading to not dying from cervical cancer. No cure means dying at the same time and from the same cause as in a situation without screening.
It is assumed that high risk is associated with non-attendance (Berget 1979; Boyes et al. 1982). This is quantified as 10% of the population never taking part in screening and having a 3 times higher risk than those who might attend to screening when invited. Positive screening results are assumed to be followed by colposcopically guided biopsy and pathology. The effects of prevention of cervical cancer through uterus extirpation for other reasons than cervical cancer will not be taken into account.

costs

Only the following medical costs have been included in this analysis: cost of pap smear screenings, visual examinations and diagnostic follow up. Time and travel costs for participating women were not included. The costs of treatment of cervical cancer and its precursors were not considered because they were assumed to be similar for both screening modalities, and the same applies for the cost of organising the screening programme.
Unit costs of Papsmear taking and examination, visual examination and colposcopy were calculated analogous to cost estimates for the Dutch situation, with use of unpublished data from the Institute of Cytology and Preventive Oncology in New Delhi.
Unit costs of biopsy were calculated relative to pap smear costs according to the Dutch situation where these costs are three times higher than for pap smear examination. A pathologist is assumed to devote 50% of his time to pap smear examination, and 50% to other related activities.
It is assumed that organisation (transport, education and publicity) is similar for pap smear and for visual examination, as are the associated costs.

results

When the population of Bangalore is offered cervical cancer screening at one age during life time, the most efficient age for this is expected to be between 40 and 50 when the prevalence of screen-detectable progressive CIN and preclinical invasive cancer is 1.4%. Figure 7.1 shows the undiscounted expected number of life years gained due to screening if one pap-smear or visual examination would be offered at different ages. The optimal age for pap-smears is about 10 years earlier than maximal incidence of cervical cancer because of the quite long screendetectable preclinical period. The optimal age for visual examination is slightly higher than for pap-smears because this screening is less sensitive early in the screendetectable preclinical period. With a yearly discount rate of 3%, the optimal age is some two years higher than without discounting.

table 7.1 Estimated costs for relevant medical procedures

	Cost elements	Number /day	Unit costs
Papsmear taking	Health worker/auxillary nurse midwife: 270 Rs/day	50	5.40
	Admin.personnel: 270 Rs./day	50	5.40
	Material costs		1.00
	Total costs		11.80
Papsmear examination	Lab. technician: 420 Rs./day	42	10.08
	Lab. assistant: 270 Rs./day	208	1.08
	Cytopathologist: 1260 Rs./day	208*	5.04
	Education: 10%		1.50
	Materials, microscope, autostainer:		4.80
	Housing:		3.00
	Total costs		25.50
Visual examination	Health worker/ auxiliary nurse midwife: 270 Rs/day	50	5.40
	Admin.personnel: 270 Rs./day	50	5.40
	Total costs		11.80
Positive screening outcome	Colposcopy		200.00
	Biopsy and histologic examination		76.50
	Total costs		276.50

** 10% of the smears are assumed to be reviewed by a pathologist; therefore 180 screens per day is here 18 examinations per day. Half of the work time of a cytopathologist is assumed to be spent on training, management activities etc.*

Table 7.2 Main effects and costs in Rs to be expected per 100,000 at start of screening in the Bangalore population from offering pap smears at age 45 compared with offering visual examination at ages 39 and 51 or at ages 37, 45 and 53 when 75% of invited women who are at risk for cervical cancer attend to screening, without discounting and with a 3% yearly discount rate

	no discounting			3% yearly discount rate		
	1 pap smear	2 visual examinations	3 visual examinations	1 pap smear	2 visual examinations	3 visual examinations
screenings	37,291	71,781	107,669	24,538	46,890	70,437
cost of screening	1,390,947	847,019	1,270,489	915,260	553,298	831,156
true positives	367	378	516	241	248	342
false positives	732	7,051	10,679	482	4,603	6,993
cost of diagnosis	303,847	2,054,034	3,095,369	200,027	1,341,294	2,028,291
prevented mortality	154	153	207	71	72	98
life years gained	2,720	2,808	3,808	954	988	1,342
C/E in Rs per death prevented	11,024	18,912	21,056	15,731	26,141	29,095
C/E in Rs per life year gained	623	1,033	1,147	1,169	1,917	2,131

Figure 7.1 also shows that a single pap-smear is expected to be almost twice as effective as visual examinations. Thus, from the perspective of effectiveness, offering pap-smears at one age could best be compared with offering visual examination at two ages. The optimal policy for offering visual examinations at two ages is determined in the same way as shown for one screening in figure 7.1, but now by varying two ages. The optimal policy is expected to be to offer the screening twelve years apart at ages 39 and 51. The cost of a visual examination is about three times lower than that of a pap smear, therefore we have also estimated the effectiveness of visual examinations at three ages where the optimal screening policy is to offer screening at ages 37, 45 and 53.

If 75% of women are attending the screening programme, the screening policies considered in this article are expected to lead to a gradual decrease of cervical cancer mortality to a maximum of 24% reduction in the total female population of Bangalore for the single pap smear policy and the policy of two visual examination and to 32% reduction for the policy of three visual examinations.

Table 7.2 compares the expected cumulative effects from one time pap-smear screening as well as two and three times visual examination. Of course the numbers of invitations and screenings needed for the policy with visual examination is about two respectively three times higher than for the policy with pap smears. The numbers of true positives, deaths prevented and life years gained are very similar for both policies, but because of the relatively low specificity of visual examination, the number of false positive screening outcomes which need further diagnostics is much higher than for pap smears.

sensitivity analysis

For answering the question under which assumptions two visual examinations at ages 39 and 51 are more cost-effective than one pap smear at age 45, parameters concerning visual examinations are varied while the parameters concerning pap smears are left constant. The sensitivity of visual examination for dysplasia and CIS is varied from 10% to 80% and for invasive cancer from 10% to 85%; that variation is represented in the figures as from 10% to 100% of the sensitivity of pap smears. Specificity of visual examination is varied from 80% to 100%.

Figure 7.2 shows for each combination of sensitivity and specificity at what cost of visual examinations two times this screening test is equally cost-effective as a one time pap smear, given the other point estimates. At the point estimates of 50% sensitivity and 90% specificity of visual examination, the pap smear test is always more cost-effective than visual examination, because of the high costs caused by positive screening results of visual examination.

Figure 7.3 shows for each combination of sensitivity and specificity at what cost of a positive screening outcome (that is: colposcopy plus pathology), visual examination is equally cost-effective as pap smears, given the other point estimates. If sensitivity of the visual examination is less than 29% that of the pap smear, then the cost of positive screening outcomes can't be low enough for visual examination being more cost-effective than pap smear taking. If sensitivity of visual examination would be as high as that of pap smears, it is only more cost-effective at the baseline cost estimates if specificity is higher than 90%.

discussion

The baseline model assumptions show that one stage screening by visual examination can be used as an effective mode of screening for the prevention and early detection of cervical cancer. The direct screening costs to be made for achieving the same effectiveness as with pap smears, are expected to be lower for visual examination because both costs of one visual examination test are estimated to be about one thirst of one pap smear test while the effectiveness of one visual examination test is estimated to be about half that of one pap smear test. Considering this, one may prefer visual examination, also because of its advantage of an instantaneous screening outcome which may prevent problems of tracing back women with positive pap smear outcomes for a follow up of further diagnostics.

If also medical costs after the screening test are included in the analysis, a very different picture appears: Due to the relatively unfavourable specificity of visual examinations, the number of diagnostic follow up procedures needed is by far higher than for pap smear screening. The follow up procedure assumed for this analysis is colposcopically guided biopsy, a much more costly procedure than the screening test itself that needs to be performed in more than 10% of women who receive a visual examination. That's why the expected costs of follow up procedures dominate the medical cost due to a screening programme by visual examinations.

The argument that screening by visual examination does not need a lot of input of technology seems to neglect that a screening programme which uses this test would lead to large numbers of diagnostic procedures such as pathological examinations that would also require the availability of extensive pathology laboratory facilities.

The difference in sensitivity between visual examination and pap smears is less for invasive cancer than for pre-invasive lesions, therefore more of the effectiveness of visual examination is due to early detection of cancer than to prevention of

Figure 7.1 Expected number of screenings and life years gained when offering one screening. The numbers in the graph lines represent the age at which screening is offered

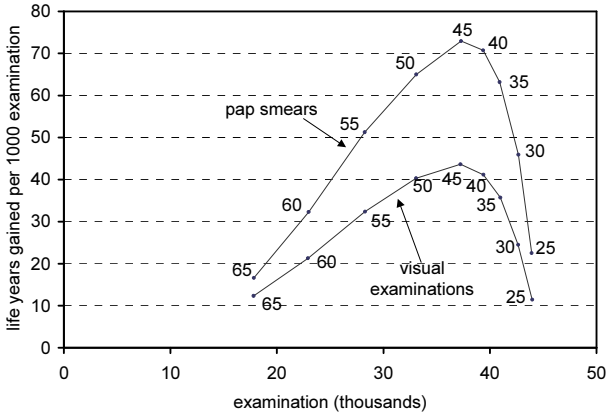


Figure 7.2 Cost of visual examination at which a policy of two visual examinations is equally cost-effective to a policy of taking one pap smear

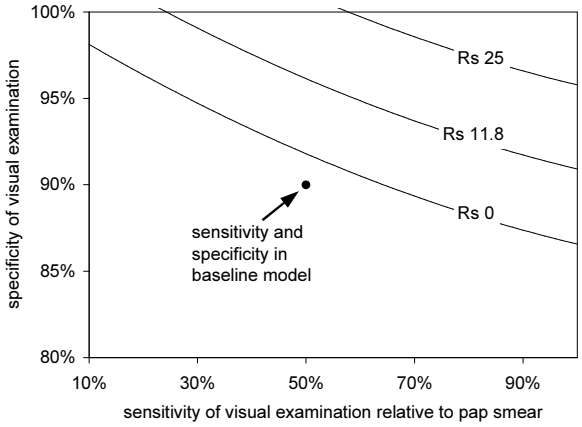
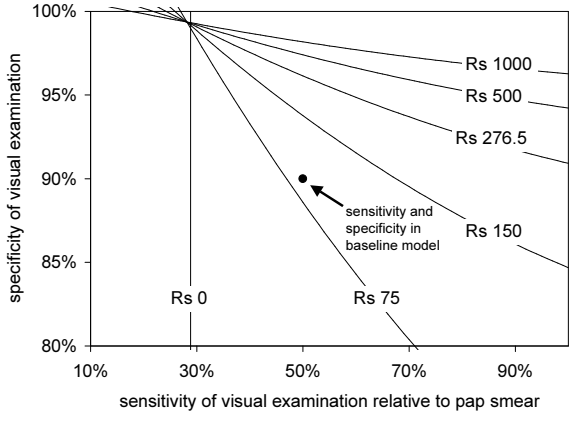


Figure 7.3 Cost of diagnostic follow up after a positive screening outcome at which a policy of two visual examinations is equally cost-effective to a policy of taking one pap smear



cancer by treatment of pre-invasive lesions. This implies that costs of treatment of invasive cancer involved with a screening programme are expected to be higher when using visual examination. These costs are not taken into account in this analysis, which gives a bias in favour of visual examination. The sensitivity analysis shows that model parameters need to be chosen considerably more favourable than the point estimates before visual examination would be more cost-effective than pap smears.
This analysis is about the application of visual examination with a high threshold for positivity. It is also possible to consider a two stage screening strategy which involves visual examination as a selection test for applying a pap smear. Then a positive outcome visual examination with a low threshold is suggested to be followed by a pap smear. However, visual examination with a low threshold is expected to lead to some 70% positive outcomes which would need a pap smear (Singh et al. 1992), therefore the estimated cost of one screening would be 11.80 Rs for a visual examination plus 70% of 37.30 Rs for a pap smear (total 37.91 Rs) which is more than the cost of one pap smear, while it is less effective because of missed cases in the 30% found negative at visual examination. Therefore already without applying our simulation model, we know that this screening strategy is almost as costly and less effective than pap smear taking, therefore it is not preferable from a cost-effectiveness perspective. Increasing the threshold for positivity of the visual examination in a two stage screening setting would improve cost-effectiveness.
There is also a possibility to screening for cervical cancer by inspection with some kind of aid, like applying solution of acetic acid to the cervix, by using a gynoscope or by cervicography (Baldauf et al. 1997; De Sutter et al. 1998; Frisch et al. 1994; Megevand et al. 1996; Van Le et al. 1993). Some of the reports on these methods look promising but so far there is a lack of published results to estimate test characteristics in a screening population.
In conclusion: Unaided visual examination does not seem to be more cost-effective than pap smear taking in the setting of a developing country. Technological input needed may also be high due to the large amount of follow up procedures.

additional discussion

During the course of the research reported in this chapter evidence was published showing that some simple aids to visual inspection are expected to lead to very little extra costs while both sensitivity and specificity improve.(Gaffikin et al. 1999; Sankaranarayanan et al. 1999; Sankaranarayanan et al. 1998) This means that unaided visual inspection of the cervix would never be the preferred screening method. Aided visual inspection has been shown to possibly have a sensitivity that is similar to pap smears. Specificity is still not very high, but so much higher than for unaided visual inspection, that it is not clear on beforehand whether the cost saved by screening by aided visual inspection instead of pap smears is superseded by the extra cost of diagnostics due to lower specificity. The expected effectiveness and need for scarce resources of visual inspection with application of several possible simple aids in comparison with pap smear screening is subject for further research.

8. IN SEARCH OF THE BEST UPPER AGE LIMIT FOR BREAST CANCER SCREENING

abstract

objectives: To determine the best upper age limit for a breast cancer screening programme.

methods:

A model based study using optimistic and pessimistic assumptions concerning improvement of prognosis due to screen-detection and duration of the period of mammographic detectability, resulting in upper and lower limits for favourable and unfavourable effects.

results:

Under pessimistic assumptions, the balance between positive and negative effects of screening remains favourable up to an age of around 80. Under optimistic assumptions, this balance does never become clearly negative with increase of the upper age limit of a screening programme.

When including the costs in the analysis, the balance between effects and costs of increasing the upper age limit from 69 to 75 is likely to be at least as favourable as intensifying a screening programme within the age group 50-69. A further increase leads to a markedly less favourable balance.

conclusions:

Competing causes of death do not lead to missing net benefit for women up to at least age 80, but the disproportional rise of negative effects of screening with age in older women, leads to a lower cost-effectiveness ratio than intensifying screening at ages 50-69.

introduction

Setting upper and lower age limits for a mass screening programme for breast cancer must be very well justified as breast cancer is an important disease in younger as well as older women. In the U.K. programme women are invited for screening every three years between age 50 and 65 (Chamberlain et al. 1993). In the Dutch programme women are invited for screening every other year between 50 and 70 years of age (de Koning et al. 1995b). In the US most institutions which give advice concerning breast cancer screening do not give an upper age boundary (Costanza 1992).

The lower boundary of age 50 has been justified by the published results of randomised trials which show a much lower (and statistically not significant) reduction in breast cancer mortality for those who entered the trial under the age of 50 as compared to women who entered at higher ages (Nyström et al. 1993). This difference in effectiveness can be explained by physical changes which occur around the menopause (like radiographic density of the breast and tumour growth rate) and which influence detection rates and earliness of detection by mammography.

There is no such clear-cut justification of an upper boundary of the invitation scheme. The relevant biology for this decision is complicated: Breast cancer mortality rates are higher at older age. Therefore screening of an older population can lead to more prevented breast cancer deaths than screening of a younger population. However, the number of life-years gained by preventing a breast cancer death by screening rapidly decreases with age of detection, because life expectancy decreases with age. Nevertheless, in the Netherlands life expectancy of a 70 year old woman is still 15 years (CBS 1992). The increasing rates for mortality from other causes than breast cancer lead to a higher probability of women dying from other causes in the lead-time period. As a consequence the number of extra incident cases generated by screening relative to the number of prevented breast cancer deaths increases with age at screening. This unfavourable effect of screening also increases with longer lead time, which is roughly proportional to the duration of the preclinical period of mammographic detectability. So, on the one hand, an equal amount of screenings at higher ages will lead to more prevented breast cancer deaths, while on the other hand, there is a less favourable balance between the positive and negative effects of screening.

The latest publication on the Swedish randomised trials (Nyström et al. 1993) reports a point estimate for mortality reduction in the age group 70-74 which is much lower than for the age group 50-69. Since there is no biological explanation for such a difference, and the number of cases involved is very small, it seems reasonable to explain this difference by random fluctuation. However, the present lack of another explanation should not lead to neglecting the possibility that indeed breast cancer screening leads to a smaller improvement of prognosis over age 70 (Fletcher et al. 1993b).

This paper attempts to determine the best upper age limit for breast cancer screening, once the decision to organize a programme has been taken. In the first place the balance between favourable and unfavourable effects is considered, but also the balance between costs and effects.

There are not many data available about the effects of screening for breast cancer in women older than 70 years. It is therefore not possible to give a precise assessment of what is to be expected from mass screening in older women. However, the determinants of such an upper age limit are identified and on basis of the existing uncertainty concerning these determinants, margins of possible outcomes are set.

table 8.1 Ratio of detection rate at prevalence screening and incidence rate in situation without screening of different screening projects

age class	K/O, incidence control group	K/O, incidence cancer registry	BCDDP	Nijmegen trial	simulated optimistic variant	simulated pessimistic variant
40-44	1.9	1.8	3.0	1.0		
45-49			3.8	2.5		
50-54	3.0	3.5	3.3	2.8	2.0	2.0
55-59			4.0	2.4	2.3	2.3
60-64	3.8	4.7	3.6	2.5	2.8	2.8
65-69			4.7	7.2	3.4	3.4
70-74	5.4	4.4		3.0	3.6	4.6
75-79	-	-	-	4.0	3.9	6.0
80-84	-	-	-	3.4	4.0	8.4

Sources: Tabár et al. 1988 and BCDDP 1979

methods

The MISCAN model for underlying natural history of breast cancer and the impact of screening, which has been used for this analysis, has been described elsewhere (van Oortmarssen et al. 1990b). The disease model is based on a 3-stage division of the development of invasive breast cancer in which the stage reflects tumour size. A proportion of the invasive breast cancers is preceded by a screen-detectable ductal carcinoma in situ (dCIS). The screendetectable stages have an exponentially distributed sojourn time with an age-dependent mean. Most model assumptions are identical to those used for the Dutch cost-effectiveness analysis which focused much on screening between the ages 50 and 70. This model reflects estimates of health effects and social costs of the primary process of screening, changes in diagnostic procedures, primary therapies, follow up after treatment, metastatic disease, terminal illness and breast cancer mortality when a two yearly screening programme is carried out during a period of 27 years, after which time the maximum impact of screening on mortality is reached (de Koning et al. 1991). The assumptions which are specifically relevant for screening in older women have been studied in more detail by relating them to the results of trials which included women over age 70: the study in Kopparberg/Östergötland (Tabár et al. 1988), the Dutch pilot project of Nijmegen (unpublished data) and the BCDDP (BCDDP 1979).

model variants

Two main variants were used:
The 'optimistic variant' which assumes no further increase in preclinical duration after the age of 65.
The 'pessimistic variant' assumes a further increase in preclinical duration with age which is extrapolated from the trend in younger age groups.
The model results have been integrated in one outcome measure: 5% discounted quality adjusted life years (QALYs) gained. The number of QALYs gained by a screening programme is calculated by attributing utilities to each health state that is relevant for breast cancer screening and by multiplying these utilities with the number of life years in each of these health states. The difference between the total number of QALYs in a situation with a screening programme and the total number of QALYs in the situation without screening is the number of QALYs gained (de Haes et al. 1991). In this way not only the effects of screening on mortality are taken into account, but also the impact of diagnostic procedures, primary therapy, follow up after treatment, metastatic disease and terminal illness.
For demonstration purposes, an attendance rate of 100% is assumed. This gives the sharpest contrast between different variants.
For calculating cost-effectiveness ratios, an attendance pattern has been assumed as it occurred in the Nijmegen trial (unpublished data). A realistic attendance pattern has been assumed because here an estimate of social costs is made, which is strongly influenced by attendance. These attendance rates decline from 75% at age 51, to 61% at age 71 and 21% at age 81. The assumptions about social costs are described in detail by de Koning et al (de Koning et al. 1991) using 1990 as base year for discounting and start of building up of the programme. The applied exchange rate is 2.7 Dutch guilders of 1990 per pound.

results

preclinical period

A measure for the duration of the preclinical period of mammographic detectability is the ratio between detection rates at first screening and incidence rates in the situation without screening. Table 8.1 shows the comparison of results from screening projects which have enrolled women over 70 and simulated results.
It is clear that preclinical duration strongly increases between ages 50 and 70. Comparison of the screening results from the Kopparberg/Östergötland trial with the incidence in the control group seems to show a further increase in preclinical duration at higher ages (Tabár et al. 1988). However a comparison with incidence from the Swedish cancer registry suggests a stabilizing preclinical duration after age 60. The results from the Nijmegen trial and

the BCDDP show at most a slow increase of preclinical duration at ages over 70.

two main variants

Table 8.2 shows the main positive effect of screening: life years gained, and two important negative effects: life-years in lead time (that is time with knowledge of the disease outside the gained life-years) and extra incidence caused by screening. These are the expected results of a screening programme which starts at age 51 and goes on screening with two year intervals up to the age in the left column. From around age 90 further extension of the screening programme leads to so few extra life years gained, that it fades away in the random fluctuation of the model outcomes. It also shows that the expected number of life-years gained is not very different in the optimistic and pessimistic variant. The important difference between the two variants is the amount of negative effects.

With extension of a screening programme to higher ages, the number of extra years in lead time and the number of extra incident cases increases much more rapidly then the number of life-years gained. The pessimistic variant shows that with extending the screening programme from age 79 to age 81, gaining 1 extra life year in the population coincides with 1 extra breast cancer case (leading to 1 extra primary treatment) and leads to more than 10 extra life years with knowledge of the disease outside the gained life-years.

Figure 8.1 shows the number of 5% discounted QALYs gained as a function of the upper age limit of invitation for screening of a programme with 2 year screening intervals, starting at age 51. There is a wide range between the optimistic and the pessimistic variant. The optimistic variant shows no clear decrease in the expected number of QALYs gained at any increase of the upper age of the screening programme.

In the pessimistic variant, the balance of favourable and unfavourable effects (measured in 5% discounted QALYs) becomes negative if screening is continued beyond age 80.

Therefore it is very likely that when extending a mass screening project up to around age 80 or even higher the balance between positive and negative effects remains favourable.

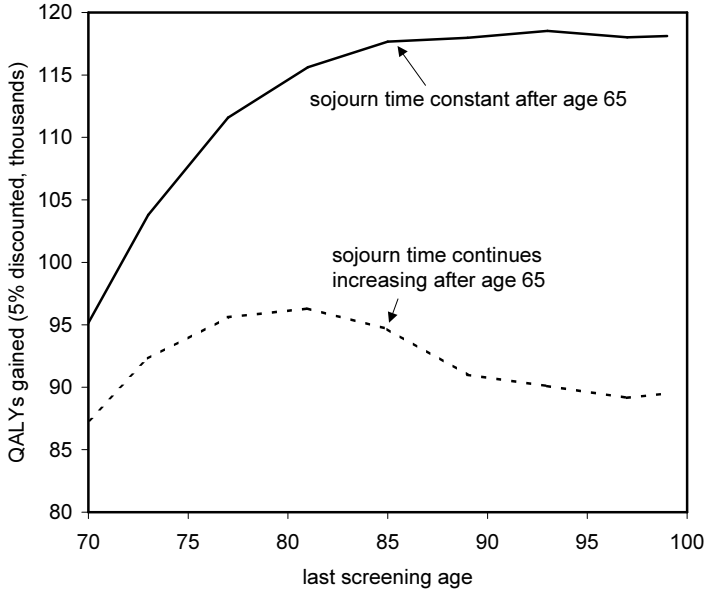
table 8.2 Life-years gained, life-years in lead time and extra number of incidents of breast cancer to be expected from a screening programme with 100% attendance and invitations every two years starting at age 51; results for the optimistic and pessimistic variants

upper age limit	life-years gained (*1000)		life-years in lead time (*1000)		extra incidence (cases)	
	opt.	pes.	opt.	pes.	opt.	pes.
69	408	395	357	377	5266	6258
71	428	416	396	433	6593	8304
73	448	436	433	495	7946	11089
75	463	450	468	560	9406	14633
77	475	466	495	632	10831	19437
79	480	476	517	703	12205	25142
81	486	483	538	776	13864	32356
83	491	489	556	848	15582	40647
85	494	492	570	909	17165	49273
87	495	495	580	973	18547	59816
89	494	496	586	1024	19656	69392
99	497	497	597	1057	22288	77760

table 8.3 Life-years gained, life-years in lead time, extra number of incident cases of breast cancer and 5% discounted QALYs gained, to be expected from a screening programme with a realistic attendance and invitations starting at age 51; results for pessimistic variant without any improvement of prognosis due to screening

upper age limit	life-years gained (*1000)	life-years in lead time (*1000)	extra incidence (cases)	QALYs gained (5% disc.)
69	292	275	4651	62727
99	292	463	17663	57855

Figure 8.1 The number of 5% discounted QALYs gained as a function of the upper age limit of invitation for screening of a programme with 2 year screening intervals, starting at age 51



cost-effectiveness
The marginal cost-effectiveness ratio (CER) in the pessimistic variant of extending a programme from a last age of invitation of 69 to a last invitation at 75 is £ 8,400 per QALY gained. This is about the same ratio as results from intensifying the invitation scheme in the age group of 50 to 70 years old. Further extension to age 79 has a marginal CER of £ 36,000 per QALY gained.

no improvement of prognosis
A model which consists of the same assumptions as the pessimistic model used for cost-effectiveness analyses, but with no improvement of prognosis due to screening at ages 70+ (therefore very close to the point estimates of the Swedish randomised trials), leads to results as shown in table 8.3. As expected in such a situation, this would have detrimental effects. Screening over 70 could lead to doubling of the years in lead time and quadrupling of the extra incidence from 0.6% to 2.4%

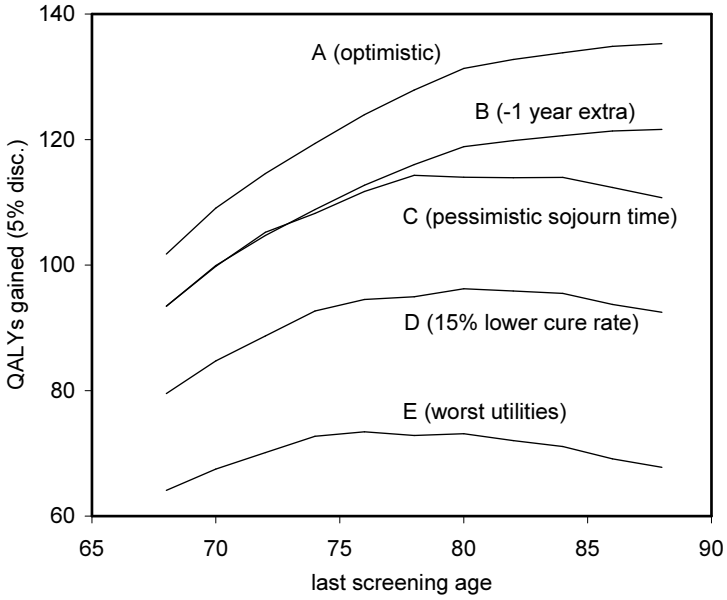
discussion
The conclusion of Forrest c.s. (Forrest 1986), (Forrest and Aitken 1990) concerning the upper boundary of screening is that because of the rapid fall in attendance after the age of 64, the invitation scheme of a screening programme should not include higher age groups. From a cost-effectiveness point of view, this seems like an overestimate of the cost of sending out invitations for screening. The cost of inviting women is relatively low as compared to the other costs, therefore a lower attendance rate should not automatically lead to an upper age limit in the invitation scheme.
Besides the upper age limit for issuing invitations, Forrest c.s. recommend to positively encourage older women to get screening without invitation; for which they do not mention an upper age limit. This recommendation does not acknowledge that from a certain age the unfavourable effects of screening outweigh the favourable effects.
Both recommendations together lead to under-serving women just over 65 as well as possibly harmful screening of very old women.
For an estimation with any precision of the costs and effects of a mass screening project for women over 70, more data on the effects of screening are needed, especially on preclinical duration of the phase of mammographic detectability. The available data leave room for a wide range of possible consequences. Nevertheless some conclusions can be drawn.
Both main variants of the model assume an improvement of prognosis due to screen-detection of breast cancer which is the same for women of all ages. We are aware of the fact that the results from the Swedish randomised trials show a mortality reduction which is far less for women in the age group of 70-74 than for women in the age groups of 50-69; i.e. 2% or 6% (depending on the way this figure is calculated) vs. 29% (Nyström et al. 1993). Only a small part of this difference can be attributed to the fact that the older women were invited for screening only two times, and that older women are less likely to attend a screening. However, there is no biological explanation for this substantial difference, so for the main variants the difference was assumed to be due to random fluctuation. The estimate of mortality reduction in the age group 70-74 is based on very small numbers of deaths, and there is no significant difference in mortality reduction between the age groups 50-69 and 70-74.
Although it seems reasonable to assume an equal improvement of prognosis due to screen-detection for women over 70 as for women from 50 to 69 years old, there is no proof of a favourable effect for women over 70. Therefore it is also necessary to consider a model which assumes no improvement of prognosis due to screening of women over 70. This assumption is very close to the point estimate from the Swedish randomised trials (Nyström et al. 1993). Such a model shows that screening women over 70 can lead to a considerable amount of unfavourable effects, which in total can cost more than 8% of the QALYs that are expected to be gained by a programme of screening women from 51 to 69 years old.
Judging only from the number of QALYs gained it would be advisable to extend breast cancer screening at least up to age 80. But, because of the relatively unfavourable balance between positive and negative effects, the number of screens that is required to obtain a certain amount of favourable effects is considerably higher than in younger age groups.
The sharp increase of cost per QALY gained with extending a screening programme to higher ages, is only for a minor part caused by the cost of screening; by far the most important cause is the strong increase in negative effects.
When only the balance between favourable and unfavourable effects is taken into account, mass screening should be continued to at least the age of 80, but when efficiency is also taken into consideration, one may be hesitant to offer screening to women older than 70 years because of an increasing cost-effectiveness ratio.
Although the results leave a wide margin for the best upper age limit, not much more improvement can be expected for the near future, since of the two possible ways to improve on the estimates, one is not likely to give much more precision, the other is not likely to be attainable.
The first is making existing data on screening of older women better available. For instance in the Kopparberg/Östergötland study data on women

invited at ages over 74 have not been published (Tabár et al. 1987). However, these data are based on small numbers, therefore it is not to be expected that better availability of data would greatly improve the precision of the estimate.
The other way is of course a new trial. Such a trial can give a better understanding of the natural history of breast cancer in older women. But if it is to serve as an investigation into the possible mortality reduction of breast cancer screening, a serious problem is encountered because of the low attendance rate to be expected in older women. Even when there is a strong effect of screening, low attendance leads to a serious dilution of the contrast between the invited and not-invited group. Such a trial therefore would need a very large population in order to reach an acceptable power.
The risk of radiation induced cancers is not included in the study. That is because of the assumption that for screening women of 50 to 70 years old this effect is small in comparison with the other effects of screening, and even smaller when screening women over 70.
Less readily quantifiable aspects of screening, such as a possible educational effect on women and the effect on quality of life after a negative screening result, have been left out of the analysis, because they are assumed to be of negligible effect on the optimal upper age limit.
This paper is based on model for the Dutch situation. This means that assumptions made about demography, epidemiology, organization and quality of the mass screening project and on costs of health care facilities are specific for the Netherlands. This does not imply that it is not possible to generalise the conclusions of this paper to other Western countries. The balance between favourable and unfavourable effects of screening can only be substantially influenced by large differences in life expectancy at the time of screening. These kind of differences do not occur among countries and regions where organizing breast cancer screening is an issue. Cost-effectiveness considerations also depend on the effectiveness of screening at ages higher than 70 relative to screening between 50 and 70 which in its turn depends on the age dependency of the risk for breast cancer. This age dependency is neither very different in Western situations.

additional discussion
This chapter concentrates strongly on the balance of favourable and unfavourable effects as a determinant for the best upper age limit for a breast cancer screening programme. The reason for it is that, at least as far as the Netherlands are concerned, the discussion on screening older women was dominated by the issue of age discrimination and there was little possibility for cost-effectiveness considerations.
Following are some additional sensitivity analyses with respect to the balance of favourable and unfavourable effects that were used for the decision on extension of the Dutch breast cancer screening programme to higher ages. After that I present some considerations on cost-effectiveness of screening older women.

further sensitivity analyses concerning the balance of favourable and unfavourable health effects of breast cancer screening at older ages
In addition to the analyses of the previous part of this chapter, the following model assumptions have been varied.
- After observing that there is good reason to assume that the Two Counties Study has shown a substantially better screening performance than so far in the Dutch national screening programme in terms of detection of particularly smaller cancers(Fracheboud et al. 1997), it was considered that there is also reason to assume that the influence of the Dutch programme on mortality is less substantial than in the Swedish randomised trials(Nyström et al. 1993). Therefore next to the basic scenarios, also a 15% lower improvement of prognosis was assumed than originally estimated from the Swedish trials.
- The basic scenarios assume that the utilities of life years spent in the health states that are possibly affected by screening, are according to the point estimates from our quality of life study.(de Haes et al. 1991) As a variant, a worst case scenario based on uncertainty of the utility estimates was created. Instead of the point estimates, the side of the range of the individual estimates was chosen that gives the model prediction of screening effects that is least favourable for screening.
- It has been observed in the Nijmegen pilot project that mortality from other causes than breast cancer is lower among those participating to screening than among other women of the same age in the same population. This difference is very large just after screening, which is clearly due to death often being preceded by a period of bad health that precludes participation to screening. But also up to several years after screening, there is still a substantial difference in mortality. We have estimated that, apart from the risk of dying from breast cancer, women of 65 years and older who participate in the Nijmegen breast cancer screening pilot project, have around 1 year more life expectancy than the average population of the same age in that project. Therefore we added a model variant that assumed that per breast cancer death prevented there was 1 more life year gained than in the basic scenario that did not assume any association between participation to screening and mortality from other causes than breast cancer.

Figure 8.2 The number of 5% discounted QALYs gained as a function of the upper age limit of invitation for screening of a programme with 2 year screening intervals, starting at age 51



Line A shows the expected number of QALYs gained under optimistic assumptions on sojourn time, 1 year extra life expectancy of screening participants, mortality reduction as estimated from the Swedish randomised trials, and the point estimates for utilities of health states that are influenced by screening.
Line B is as line A but without the 1 year extra life expectancy of screening participants.
Line C is as line B but with pessimistic assumptions on sojourn time.
Line D is as line C but with 15% less probability of cure due to screen-detection than estimated from the Swedish trials.
Line E is as line D but with the worst case scenario on utility estimates.

table 8.4 Simplified utilities

health state	period of life	utility
advanced breast cancer	1.768 years prior to breast cancer death	0.624
primary therapy	1 year from diagnosis of breast cancer	0.75
disease free period	from end of period of primary care to advance breast cancer or death from other causes	0.955

table 8.5 Simplified cost assumptions

	cost in Euro
screening	35
diagnosis	0
primary and adjuvant therapy	7,000
advanced breast cancer	20,000

Figure 8.2 shows the expected number of 5% discounted QALYs gained of different scenarios by the number of two-yearly screenings that is added after the screening programme of women aged 50-69.
Figure 8.2 shows that optimistic scenarios for screening have no clear upper boundary for the age from which the unfavourable health effects of breast cancer screening outweigh the favourable effects. The decreasing slopes by age are not only caused by a decrease of the number of QALYs gained by screening, but also to a major extent by the decrease of the population with increasing age. Under the most pessimistic scenario, there remains a positive balance of favourable and unfavourable health effects up to around the age of 75.
It was considered to be prudent to extend the age range of the Dutch national breast cancer screening programme with three invitations from an age range of 50-69 to 50-75. Evaluation of the findings from screening women aged 70-75 is expected to give a better estimate of the age dependency of the preclinical screen-detectable period, thus answering the question whether reality is closer to the optimistic or to the pessimistic duration scenario. In case reality appears to be closer to the optimistic scenario, and if the same line of reasoning would be continued, then further extension of the screening programme would be appropriate.

efficiency of breast cancer screening at older ages

The line of reasoning by which screening is continued up to the age at which the unfavourable effects start outweighing the favourable effects is questionable. The cost per QALY gained of screenings that are close to the age at which no QALYs are gained, is very high because close to that age there are already very few QALYs being gained. Therefore, from a perspective of maximising utility gained given a certain amount of expenses, decision making would be based on a different line of reasoning. Then the question to be answered is: How are the invitations of the screening programme to be scheduled in order to achieve the most QALYs gained at a certain cost?
The model that has been developed for the

Dutch screening programme was used to calculate expected health effects of a range of possible programmes. Two model variants with and without further increasing preclinical duration of the disease after age 65 (resp. pessimistic and optimistic, and line C and B of figure 8.2) were considered. Invitation schedules that include women younger than 50 were not considered in order to avoid the discussion of effectiveness of screening among younger women and because such schedules were expected to be of no influence on choosing the upper age limit. For estimating cost/utility ratios, simplified models for costs and for the influence of health status on utility were applied, see tables 8.4 and 8.5
Table 8.6 shows the outcomes of this simplified model for all invitation schedules that have been evaluated. Within the list of invitation schedules, some appear to be dominated by another in the sense that the other schedule is expected to yield more QALYs at a lower or equal cost. Because of the granular character of differences between invitation schedules, there is also a possibility of extended dominance, meaning that offering a fraction of the population one schedule, while the rest will be offered another schedule, is expected to lead to more QALYs gained at lower or the same cost than the schedule that is dominated. Invitation schedules that are not dominated by others are considered to be efficient. There can be several efficient schedules because this definition of efficiency does not include a preference considering the level of the cost-effectiveness ratio.(Cantor 1994) The original invitation schedule of the Dutch programme is not efficient under optimistic nor pessimistic assumptions. Neighbouring efficient schedules under optimistic assumptions are: 50(2.5)72.5 and 50(2.5)75, and under pessimistic assumptions: 50(2.5)72.5 and 50(2)72, all preferred alternatives have a higher upper age limit for the invitation schedule. Under optimistic assumptions the original Dutch invitation schedule (50(2)68) is expected to gain 4.2% fewer QALYs then efficient screening at the same cost, and under pessimistic assumptions that is 1.8%.
Both under optimistic and pessimistic assumptions, the invitation schedule 50(2)76 is efficient. The current Dutch invitation schedule of 50(2)74 is efficient under pessimistic assumptions, but dominated by 50(2.5)75 and 50(2)76 under optimistic assumptions. But even under optimistic assumptions, efficient screening at the same cost is expected to lead to only 0.2% more QALYS gained than the current invitation schedule.
In conclusion: Even though decision making was based only on considerations about the balance between favourable and unfavourable effects in order to warrant some certainty about this balance being positive, the present invitation schedule is efficient or close to efficient when point estimates are assumed for all parameters other than average sojourn time by age.

table 8.6 QALYs gained and total costs (in millions of Euro) per million women alive at start screening programme and cost-effectiveness ratio (in Euro per QALY gained), all 5% discounted

	optimistic model			pessimistic model		
invitation schedule*	QALYs gained	costs [mln euro]	cost/QALY gained	QALYs gained	costs [mln euro]	cost/QALY gained
50(2)66	18,577	60.28	3,245	18,473	130.64	7,072
50(2)68	20,029	65.51	3,271	19,923	143.09	7,182
50(2)70	21,292	70.64	3,318	21,128	155.98	7,382
50(2)72	22,365	75.50	3,376	22,170	169.10	7,627
50(2)74	23,218	80.26	3,457	22,967	182.89	7,963
50(2)76	23,993	84.76	3,533	23,607	197.91	8,384
50(2)78	24,645	89.12	3,616	24,024	214.79	8,941
50(2)80	25,109	93.52	3,725	24,159	233.68	9,673
50(2)82	25,427	97.59	3,838	24,083	254.29	10,559
50(2)84	25,620	101.32	3,955	23,864	274.73	11,512
50(2)86	25,727	104.66	4,068	23,529	296.24	12,590
50(2)88	25,772	107.53	4,172	23,124	317.95	13,750
52(2)68	17,985	56.78	3,157	17,842	124.40	6,972
52(2)70	19,243	61.91	3,217	19,032	137.32	7,215
54(2)70	17,041	53.96	3,167	16,757	120.57	7,195
54(2)72	18,118	58.81	3,246	17,804	133.67	7,508
56(2)74	16,776	56.16	3,348	16,388	131.63	8,032
58(2)76	15,340	53.85	3,511	14,927	131.62	8,817
60(2)78	13,887	51.76	3,727	13,180	134.79	10,227
50(2.5)67.5	17,444	51.34	2,943	17,421	112.01	6,430
50(2.5)70	18,877	56.25	2,980	18,784	124.73	6,640
50(2.5)72.5	20,019	60.98	3,046	19,912	138.11	6,936
50(2.5)75	20,895	65.63	3,141	20,673	152.99	7,400
50(2.5)77.5	21,688	69.93	3,224	21,229	170.41	8,027
50(2.5)80	22,223	74.36	3,346	21,359	191.10	8,947
50(2.5)82.5	22,537	78.44	3,481	21,219	213.97	10,084
50(2.5)85	22,688	82.15	3,621	20,899	236.82	11,332
50(2.5)87.5	22,748	85.35	3,752	20,426	262.13	12,833
52.5(2.5)70	16,687	47.63	2,855	16,430	106.74	6,497
55(2.5)72.5	15,397	44.84	2,912	15,106	104.07	6,889
57.5(2.5)75	13,892	42.61	3,067	13,628	103.66	7,607
60(2.5)77.5	12,334	40.64	3,295	11,700	108.04	9,234
50(1.5)63.5	18,851	71.62	3,799	18,732	154.53	8,250
50(1.5)65	20,110	77.79	3,868	19,915	168.20	8,446
50(1.5)66.5	21,423	83.53	3,899	21,291	180.83	8,494
50(1.5)68	22,665	89.01	3,927	22,485	193.76	8,617
51.5(1.5)65	18,314	68.80	3,757	18,155	148.66	8,188
53(1.5)66.5	17,713	66.10	3,732	17,696	143.03	8,082
54.5(1.5)68	17,085	63.54	3,719	16,780	139.53	8,315
56(1.5)69.5	16,315	61.28	3,756	16,040	135.90	8,473

* given are: first age (interval) last age of the invitation schedule

9. A LONGER BREAST CANCER SCREENING INTERVAL FOR WOMEN OLDER THAN 65 YEARS?

abstract

background

The observed increase of sojourn time of preclinical breast cancer raises the question whether women aged 65 and over may be screened less frequently than younger women.

methods

A cost-utility analysis using a computer model that simulates demography, epidemiology and natural history of breast cancer to estimate expected life-years gained, extra incidence, extra life years with disease and costs as induced by different breast cancer screening programs in the general population

results

The estimated ratio of favorable/unfavorable effects is lower for longer screening intervals than for shorter intervals. The cost-effectiveness ratio is much less favorable in shorter intervals.

conclusions

A longer sojourn time of preclinical cancer should not necessarily be accompanied by a longer screening interval, but a short interval is not very efficient.

background

Now that the US government has decided to stimulate breast cancer screening among women aged 65 and up by providing Medicare coverage for yearly mammography screening (Eastman 1997),(van Oortmarssen et al. 1990b), the next step is to determine what is the best screening interval for this group.

The screen detectable preclinical period of breast cancer increases with age, as shown in an earlier model-based study on the HIP trial.(van Oortmarssen et al. 1990a). This is at least partly due to the growth rate of preclinical cancer decreasing with age (Peer et al. 1993). The question is therefore whether it would be appropriate to apply a longer screening interval for women over the age of 65 than for those aged 50-64. In the past, we noted that the balance between favorable and unfavorable effects of breast cancer screening in women over 70 is a very intricate problem (Boer et al. 1995a). We think a longer screening interval does not necessarily favorably affect this balance, as will be shown in this paper.

unfavorable effects

The two main unfavorable effects on the quality of life of screened women are extra incidence (defined as: detection of cases of breast cancer which would not have been diagnosed without screening because the woman would have died from other causes before clinical diagnosis) and a longer period of knowledge of and follow-up for breast cancer because of lead time. A slow-growing tumor is more likely to cause both types of unfavorable effects than a fast-growing tumor. With increasing growth rate, the probability of screen-detection of a cancer that would not be diagnosed without screening, approaches zero. The fast-growing screen-detected tumor has greater chance that clinical diagnosis (due to symptoms, if there were no screening) would occur before death from other causes, and its lead time can only be short. A long screening interval will cause most of the slowly growing tumors to be detected, thereby generating most of the potentially unfavorable effects of screening. The additional screen-detected cancers yielded by reducing the screening interval from, for example, three to two years, will be relatively fast-growing tumors which would otherwise have become manifest as interval cancer in the third year after screening. These cases hardly contribute to extra incidence and have a lead time of less than three years, because very few women die in the period between two and three years after the previous screening, while their early detection may well lead to the prevention of breast cancer deaths.

model assumptions

We used the Miscan program (Habbema et al. 1985) to analyze the consequences of several screening policies.

The most relevant assumptions for this research are as follows:

The model assumes an exponential distribution of preclinical disease states and an average duration of the screen-detectable period as appearing from detection rates at first screenings divided by the incidence in a situation where no screening is performed, increasing from 2.0 years in the age group 50-54 to 3.7 years in the age group 65-69 and 8.7 years in the age group 80-84 (van Oortmarssen et al. 1990b),(de Koning et al. 1995a). The screening policies analyzed comprise a basic program of screening at intervals of two years in the group aged 50-64, possibly extended to include the age group 65-94 with screening intervals of one, two or three years.

Other assumptions in the model are: The natural history of breast cancer is modeled as a progression through a number of states. The first state is 'no breast cancer' in which women reside until a transition occurs to one of the pre-clinical states when the tumor becomes detectable by screening. There is one DCIS (Ductal Carcinoma In Situ) state and 4 invasive states in the model, according to T-categories (T1a, T1b, T1c, T2+). The duration in each of the different states follows an exponential distribution. The transition to the clinically diagnosed

table 9.1 Expected favorable and unfavorable effects from a total screening starting at age 50 and with a screening interval of 1, 2 and 3 years in the age group 65-82, compared to the basic program of screening in the age group 50-64 with an interval of 2 years

	screening interval over age 65			basic program
	1 year	2 year	3 year	
per 1000 women at start program				
screens	6200	4400	3800	2600
breast cancer deaths prevented	6.2	5.5	5.0	2.9
life years gained	83	76	71	54
extra incidence	6.5	5.8	5.4	0.6
extra life years with disease	140	125	114	41

table 9.2 Expected favorable and unfavorable effects of extending a screening program to ages 65-82 with an interval of 1, 2 and 3 years, relative to the basic program

	screening interval over age 65			basic program
	1 year	2 year	3 year	
per 1000 screens				
breast cancer deaths prevented	0.9	1.5	1.8	1.1
life years gained	8	12	15	21
extra incidence	1.7	3.0	4.2	0.2
extra life years with disease	28	48	64	15
per life year gained				
extra incidence	0.21	0.24	0.27	0.01
extra life years with disease	3.5	3.8	4.2	0.7

states (with the same subdivision) is governed by the incidence and clinical stage distribution data. In the case of early detection women will enter the screen-detected states (again with the same subdivision). The two end-states of the model are death from breast cancer and death from other causes, as based on mortality data.

Age-specific assumptions on the mean duration of the (5 preceding) screen-detectable pre-clinical states of breast cancer and the sensitivity of screening had been validated using all data from the Dutch screening projects, in Nijmegen ages 35 and over, and in Utrecht ages 40 and over, covering different periods and screening intervals (van Oortmarssen et al. 1990b), (de Koning et al. 1991). Sensitivity is state-dependent: 0.4, 0.65, 0.8, 0.9 and 0.95 (ages 50+) for respectively DCIS, T1a, T1b, T1c and T2+. The mean duration of pre-clinical screen-detectable period was approximately 1.8 years at age 35 to 6.2 years at age 70. These assumptions resulted in a good fit between model predictions and observed detection rates and interval cancers (both by age, stage, screening round and interval) in the Dutch screening projects(van Oortmarssen et al. 1990b). The estimated mean dwelling times from the validation procedure have later been slightly adjusted based on the more recent experience concerning detection rates at first screenings from the Dutch national screening program. Other epidemiological parameters such as incidence, stage distribution and mortality in the situation without screening were based on Dutch data.

Applying screening to a population causes a shift from diagnosing relatively large clinical cancers towards earlier (screen-detected) stages resulting in a decrease in breast cancer mortality, as shown in the randomized screening trials. Women with screen-detected cancers can have a reduced risk of dying from breast cancer, in the model depending on the cancer size at detection. The size of this improvement in prognosis after early detection has been estimated from the Swedish randomized trials(de Koning et al. 1995a).

balance of favorable and unfavorable effects

Table 9.1 and 9.2 show the expected effects of a screening program with 100% attendance, comparing the basic scenario of two yearly screening only in the age group 50-64, with scenarios including the additional effects of also screening in the age group 65-82 with an interval of 1, 2 and 3 years.

Table 9.1 shows the most important effects due to the whole program expressed per thousand members of the total female population alive at the start of the program. Each addition to the program will increase both favorable and unfavorable effects. The expected number of deaths prevented per thousand women increases from 2.9 to 6.2 when adding yearly screening of women over 65 to the basic scenario. Similarly the extra incidence increases from 0.6 to 6.5 per thousand women.

The first section of table 9.2 describes the expected effects expressed per 1000 screens in the basic scenario and expected extra effects per 1000 extra screens relative to the basic scenario in the scenarios with screening over age 65. 'Extra' here means that table 9.2 shows the differences in effects and in numbers of screenings between a scenario with screening over 65 and the basic scenario.

The last section of table 9.2 describes the balance of favorable and unfavorable effects by dividing the expected unfavorable effects by the life years gained from the basic scenario and the extra effects from the other scenarios relative to the basic scenario.

A screening interval of 3 years over age 65 already yields a large proportion of the maximally possible breast cancer mortality reduction. Increasing the screening frequency will therefore yield no substantial increase in the number of breast cancer deaths prevented. The expected number of life years

gained per thousand women by the three yearly scenario is already 71, that increases only to 83 for the yearly screening scenario (see table 9.1). This means that shortening the screening interval results in a lower number of life years gained per 1000 screenings: a decrease from 15 to 8 life years gained per 1000 screenings by changing the interval from 3 to 1 year.
Screening women aged 65 and up is expected to result in more unfavorable effects due to early detection (extra incidence and extra life years with disease) than screening under age 65. Increasing the screening frequency, however, leads to a less than proportional increase in negative effects: a decrease from 4.2 to 1.7 extra incidence per 1000 screenings. As these negative effects increase even less than the positive effects, a higher screening frequency yields a more favorable ratio between the two. Hence, weighing the favorable against the unfavorable health effects gives ratios for extra incidence and extra life years with disease per life year gained of 0.27 resp. 4.2 for three-yearly screening decreasing to 0.21 resp. 3.5 for yearly screening. Thus, judging the balance between favorable and unfavorable effects, there would appear to be no reason for a longer screening interval at higher ages.

balance of costs and life years gained
From the perspective of efficiency the question of whether a screening interval of 1 year is worthwhile is certainly a legitimate one. Table 9.3 presents a cost-effectiveness estimate. Highly simplified cost assumptions were used, as further precision has no relevance for this particular discussion. The cost per screening was set to U$ 100 with a net zero balance of other costs induced and saved by screening, such as diagnostics, primary therapy and (prevented) palliative care. The expected cost of reducing the screening interval from 2 years to 1 year was estimated as 38 thousand dollars per extra life year gained, or 4.6 times the cost in the case of a screening interval of 2 years in the age group 50-64.

table 9.3 Roughly estimated marginal cost-effectiveness ratios with 3% yearly discount rate

screening interval over age 65	extra cost per extra life year gained in U$
3 years	9,600
2 years	19,700
1 year	38,600
basic program (average cost per life year gained)	8,300

discussion
This paper presents two findings: 1: the balance of favorable and unfavorable effects of breast cancer screening improves with increasing screening frequency, and 2: the cost per life year gained increases rapidly with increasing screening frequency. Finding 2 is not so surprising as similar results have been presented for other age ranges of breast cancer screening (de Koning et al. 1991) as well as for instance for cervical cancer screening(Eddy 1990),(van Ballegooijen et al. 1992a). We present that finding primarily as a reminder to moderate possibly too much enthusiasm for a very short screening interval. Our examination of robustness of our outcomes therefore concentrates on finding 1.
Firstly we have checked the influence of extending the age range from upper age 82 to practically all women in the Medicare system by choosing age 94 as upper limit. As to be expected, this leads to a stronger increase of unfavorable effects than of life years gained (Boer et al. 1995a). For two yearly screening the extra incidence per life year gained increases from 0.24 to 0.52 and the number of extra life years with diagnosed breast cancer per life years gained increases from 3.8 to 5.1. However the improvement of the balance of favorable and unfavorable remains: the extra incidence and number of extra life years with disease per life year gained decreases from 0.59 resp. 5.5 for three yearly screening to 0.45 resp. 4.5 for 1 yearly screening.
Another strong influence on the balance of favorable and unfavorable effects is the duration of the screendetectable preclinical period. For the age range of around 50-69 the average duration of this period has been well established due to the many screening studies that have taken place in this age range. It can be argued that the apparent increase of mean duration does not extend to older ages(Boer et al. 1995a). A model in which the mean duration of the screendetectable preclinical period does not further increase after age 65 expects much lower rates of unfavorable effects because stable rather than increasing lead times after age 65 result in a lower probability of dying from other causes than breast cancer during the period of lead time after screen detection. This effect appears most clearly when considering an upper age limit of 94. For two yearly screenings the extra incidence per life year gained decreases from 0.52 to 0.12 and the number of extra life years with disease per life years gained increases from 5.1 to 2.0. However, also in this scenario, the improvement of the balance of favorable and unfavorable remains: the extra incidence and number of extra life years with disease per life year gained decreases from 0.14 resp. 2.2 for three yearly screening to 0.11 resp. 1.9 for 1 yearly screening.
The different alternative scenarios concerning the upper age limit and the duration of the

screendetectable preclinical period have little influence on cost-effectiveness ratios. The expected cost per life year gained is slightly higher for the alternative scenarios than for the basic scenario, but cost of 3 yearly screening does not exceed U$ 10,800 per life year gained and the cost of 1 yearly screening does not exceed U$ 40,700 per life year gained. This is not intended to show that the presented cost-effectiveness ratios are highly accurate, but to show the robustness of the finding that an increasing screening frequency entails a strong increase in cost per life year gained.

We expect that our finding of an improving balance of favorable and unfavorable effects with increasing screening frequency, combined with an increasing cost per life year gained, extrapolates to all cancer screening that aims for detection of invasive disease. One example of that might be prostate cancer screening. The principal finding possibly does not apply to cervical cancer screening or to endoscopic colorectal cancer screening. In these cancers very frequent screening perhaps in the end practically only leads to more early detection of the precursors of invasive disease, which would not lead to further improvement of prognosis, while the probability of finding regressive lesions (an unfavorable effect) can still increase.

conclusion

In striving to optimize health effects, a longer sojourn time of preclinical cancer should not necessarily be accompanied by a longer screening interval. Frequent screening, however, is not likely to form an attractive option from the point of view of efficiency. Breast cancer screening in women aged 65 and up thus involves making a particularly difficult trade-off between effectiveness and efficiency.

10. IMPORTANT INFLUENCES ON EFFECTIVENESS AND COSTS TO BE CONSIDERED IN THE EVALUATION OF CANCER SCREENING

introduction

Trials for cancer screening are designed to give an answer to the question whether screening is effective, in particular if it can reduce cancer mortality. The main result of such a design is presented as a relative risk of dying from cancer of the study group as compared to the control group. This relative risk is often treated as sort of a constant which does not depend on the particular screening situation as for instance by (Elwood et al. 1993). But it has been shown that the cost-effectiveness ratio can differ strongly with economic context (Brown and Fintor 1993). Besides economic context, effectiveness and costs of screening also strongly depend on several other aspects of the screening situation in question. In this chapter we describe such aspects and try to give some examples of quantification of their influence as estimated with the aid of the MISCAN simulation package (Habbema et al. 1985; Loeve et al. 1999) with examples from mainly the models on breast cancer screening (de Koning et al. 1995a; de Koning et al. 1991; van Oortmarssen et al. 1990b), because for this type of cancer screening we have experienced most international interest. Other examples are mostly from models on cervical cancer screening (Koopmanschap et al. 1990; van Ballegooijen et al. 1992a; van Ballegooijen et al. 1995; van Ballegooijen et al. 1990; van Ballegooijen et al. 1997). Our models on prostate cancer screening (Boer et al. 1997) and colorectal cancer screening (Loeve et al. 1999) are not yet suitable for public health decision support because the evidence on efficacy is too preliminary for prostate cancer screening in general and for colorectal cancer screening if anything else but FOBT screening is to be considered, therefore no results from these models are presented in this chapter.

Though results from trials are essential for sensible public health decisions on cancer screening, in this chapter it is shown that they cannot be extrapolated to other screening situations without taking into account epidemiology, demography, screening quality, policy and history, clinical practice and costs.

This chapter can be regarded as an annotated list of possible ingredients in a cookery book for screening evaluation. This list as such, inevitably doesn't appear as a very coherent narrative, but it is essential to carefully consider your ingredients in order to come to a good meal.

epidemiology in the situation without screening

Aspects of the epidemiology of different cancers that are of major importance are on the one hand the level of risk for a cancer by age which can be expressed by two measures: incidence and survival or mortality from that cancer; on the other hand stage distribution in a situation without screening and the strongly related duration of the period in which tumours can be detected by screening. The number of cancer deaths prevented and life-years gained by screening is roughly speaking proportional with the level of risk for cancer. Because the level of risk is very different among populations for which screening is considered, it is of great importance for the cost-effectiveness of screening. For instance in our analysis of the cost-effectiveness of breast cancer screening in different countries of the European Union, we found that when applying a similar screening strategy in Spain the expected cost per life-year gained is more than twice as high as in the United Kingdom (U.K.) which is mainly caused by the fact that mortality from breast cancer in the U.K. is about twice as high as in Spain (van Ineveld et al. 1993). Table 10.1 compares the crude rate of breast cancer mortality and number of life-years gained per 1,000 screens and cost-effectiveness ratio estimated for equally similar screening programmes in Spain and in the U.K.

For cervical cancer these differences in the level of risk depending on the region where screening is considered are even much greater; the pattern of age distribution of the risk also varies strongly (Gustafsson et al. 1997).

Besides substantial regional variation in risk, there can also be great variation over time. For instance the risk for cervical cancer appears to strongly decrease by birth cohort in the Netherlands over the past decades leading to a trend of decreasing mortality from cervical cancer over time which would also occur without any screening effort. Therefore the effect from screening is to be assumed much smaller than would appear from the observed decrease in cervical cancer mortality. (van Ballegooijen 1998b).

Stage distribution is also observed to differ strongly between populations. One can safely assume that cancer usually is by and large a process in which one

table 10.1 Relation between risk level and cost-effectiveness of a breast cancer screening programme (van Ineveld et al. 1993)

	Spain	U.K.
mortality (crude rate per 100,000 women)	25	52
life-years gained per 1000 screens	8.2	19.6
cost per life-year gained (£/ly)	4900	2000

table 10.2 Association between prevalence/incidence ratio and stage distribution among clinically detected cases of breast cancer in and around Florence, Italy

	Florence District Programme	City of Florence
incidence in age group 50-69 = I (per 1000)	1.6	2.3
fraction T2+ cases diagnosed clinically before screening	61%	42%
detection rate at prevalence screen = P (per 1000)	6.4	7.4
P/I	3.9	3.2

moves in one direction from relatively favourable stages to increasingly worse stages of the disease. A worse stage distribution therefore implies that on average the disease process has been going on for longer, and that the period of possible detection by screening is longer.
A very good example of this for breast cancer screening can be found in and around the Italian city of Florence. After running the Florence District Programme in areas around Florence for several years, the same executive group started screening in the City of Florence as well. While the breast cancer incidence in the City is about one and a half times higher than in the District, table 10.2 shows that the detection rates at first screenings are not so much higher (Paci et al. 1995). One can assume that stage distribution without screening and the screen-detectable course of the disease are a Markov process in which the natural course of the disease is the same, while the time of diagnosis can be earlier or later. That would mean that transition rates for growing from one to the following preclinical stage are the same in the District and City, while the more favourable clinical stage distribution is reached because of higher transition rates to clinical detection. This assumption would predict the screen-detectable preclinical period of the City to be 0.77 of that in the District, while the observed difference in prevalence/incidence ratio is 0.82, which seems close enough to support the assumption.
A worse stage distribution in a situation with the associated longer lead times is expected to cause a larger probability of benefit for a woman taking part in screening, but also to more negative effects associated with lead time, especially a larger probability of detecting cancer which would not be detected in a situation without screening.
A worse clinical stage distribution leads to a higher detection rate at first screenings and to a somewhat larger difference between stage distribution of screen-detected cases and the stage distribution as it was before screening started. Hence a worse clinical stage distribution is expected to lead to a larger extra demand for health care facilities for the primary therapy of cancer due to screening. For instance there is to be expected an increase in the demand for radiotherapy at the start of a breast cancer screening programme and to an even stronger increase in the number of diagnoses and treatments of non-palpable tumours requiring stereotactically guided biopsies.

demography
Demography describes the size of the population, the age distribution of the population and total mortality. The major influence of demography is obvious: the size of the population is proportional with the total effects in numbers and costs of a screening programme and therefore for the logistics and financing of the programme, especially when starting it. Population size is not of great influence on efficiency (the cost-effectiveness ratio) of the programme, as there is not much scale effect. Age distribution of screenings performed can have an important influence: in general a younger population leads to finding fewer cancers, but to a larger number of life-years gained per cancer death prevented by the programme. Total mortality is also of influence since for instance a lower life expectancy at a certain age leads to fewer possible life-years gained by preventing cancer death, as well as to a higher probability of detecting a cancer at screening which would not have been diagnosed or would not have led to dying from the cancer in question in a situation without screening.
However, when comparing different geographical areas where implementing a breast cancer screening programme or advocating routine screening is being considered, demography will probably not be of distinguishing influence on cost-effectiveness since no major differences are to be expected concerning age distribution and total mortality between areas of the western world, while in non-western situations mass screening for breast cancer is not (yet) a serious option. The situation is rather different when screening for cervical cancer is considered. for that cancer most screening is performed in countries with a relatively low risk for cervical cancer, while most women at higher risk for this cancer live in third world countries that do not have extensive screening programmes. When considering to introduce cervical cancer screening more extensively in a third world country, the probably much lower life expectancy and different age distribution of the population should be taken into account. Also when considering a screening programme for another entity than a geographical area, such as employees or participants to a particular health insurance, it is very well

possible that age distribution of the screened women is of major influence on cost-effectiveness.

screening quality

Two major aspects of quality of breast cancer screening are sensitivity and specificity of the screening test. A low sensitivity can be thought of as the inability of finding smaller tumours which can be found by a screening test of better quality; in that case lead time will be shorter and the probability of benefit from the screening is less. It may also be caused by randomly missing tumours irrespective of their size. This does not lead to a shorter lead time, at least not at the first screening, therefore the average screen-detected case will experience the same benefit from a first screening as in a situation with good sensitivity, but the probability of getting this benefit is smaller.

At repeat screenings the situation becomes more complicated in both possibilities of low quality.

Low specificity does not lead to a lower probability to benefit from the screening, but to a higher probability to suffer from it. False positive screening outcomes lead to undue anxiety and the burden of diagnostic procedures which also gives rise to extra costs.

For a screening programme in Germany, we estimated the effect of an increase of 12% in sensitivity at 10% more life-years gained; while a 10% decrease in positive predictive value, is expected to lead to 3% increase in total social cost of the programme (Warmerdam et al. 1997).

screening policy

Screening policy is usually either an organised programme with an invitation scheme defined by the ages at which individuals are invited to be screened (ages at which programme starts and ends, and the intervals of invitations), or it consists of making screening available and timing depends primarily on individual decisions.

Figure 10.1 shows one aspect of screening policy which is of great importance for the effect to be expected from breast cancer screening. Mortality reduction in a screening programme strongly increases with the number of screenings offered, however the effectiveness per screening decreases The percentage of mortality reduction as presented here is much lower than mortality reduction as reported from trials since the figure here is on breast cancer mortality in the total population, while a trial considers only the invited cohort.

Figure 10.1 Expected mortality reduction from the Dutch national breast cancer screening programme when assuming 70% attendance by the number of invitations increasing from 0 to 20 issued to the age range of 50-69

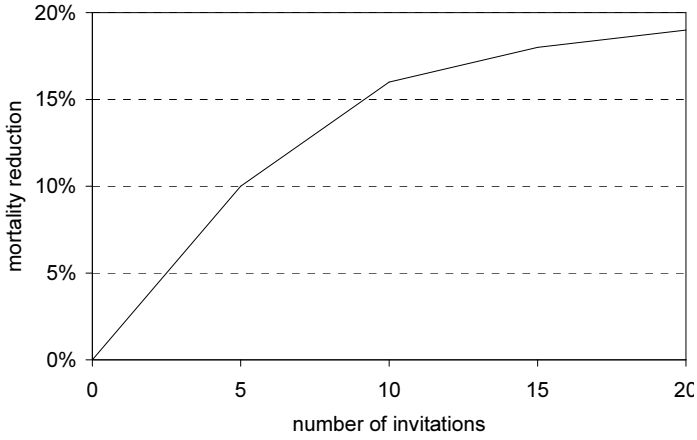
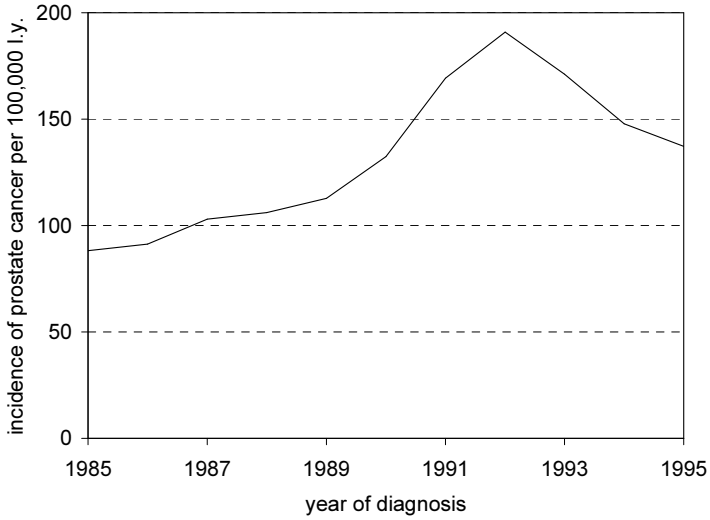


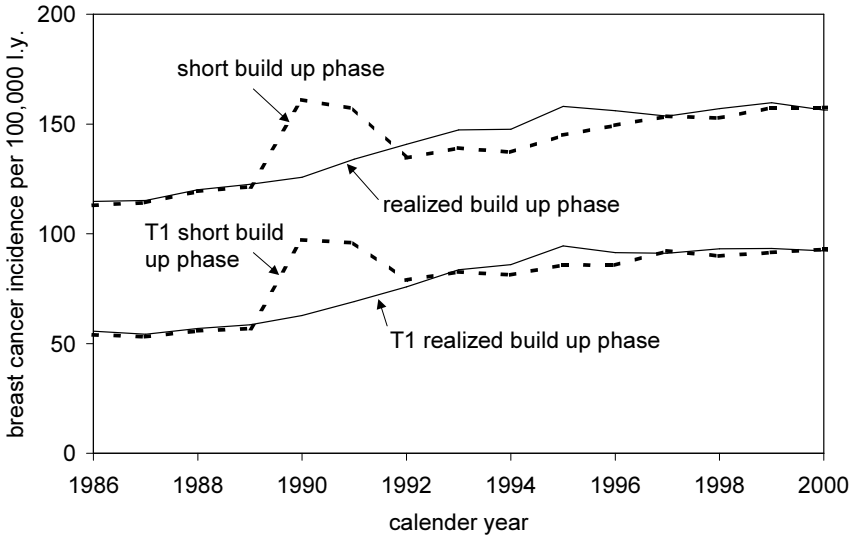
Figure 10.2 Age adjusted incidence of prostate cancer in the American SEER population by year of diagnosis from 1985 to 1995 (National Cancer Institute, 1998)



A programme screening policy and an opportunistic screening policy which would both lead to the same number of screenings are expected to show the following differences: Programme screening will lead to a more even spread of screenings over the population, more women being screened, with longer intervals; and the intervals are more evenly spread over the ages of high risk for the cancer. If these expectations are correct, a programme leads to higher effectiveness per screening than a so-called opportunistic approach.

Because detection rates at first screenings tends to be higher than at repeat screenings, the introduction of screening in a population can lead to a temporary strong increase of cancer incidence. A rather spectacular effect of that was observed for prostate

Figure 10.3 Expected incidence in the Dutch population of all breast cancer and of cases of T1 breast cancer as expected with the realised build up of the national breast screening programme and with a immediate introduction of the programme to all of the target population



cancer by the SEER registry in the US. Around the year 1992 there was a sharp peak in prostate cancer incidence (see figure 10.2). This is in agreement with introduction of PSA screening in the US population that was spread out over only a few years. Such rapid introduction would lead a sharp increase in age adjusted incidence. After the period of introduction in the population, most men only receive repeat screenings, therefore incidence would also sharply decrease again.

This doubling of new cases of cancer of course led to a major impact on the demand for health care facilities, in this case mainly prostatectomies.

In breast cancer something similar can be observed. The effect will be not as spectacular in mammographic screening because the prevalence/incidence ratio at first screenings is by far not as high as in PSA screening. Still then, around the introduction of the breast cancer screening programme in the Netherlands there was concern whether the increase in numbers of new cases of breast cancer would lead to problems with the capacity of radiotherapy, also because early detected breast cancer can often be treated by lumpectomy with radiotherapy instead of mastectomy without radiotherapy for cases which are detected not so early. Figure 10.3 shows that the gradual introduction of the national screening programme led to an also gradual increase in the demand for radiotherapy, while an instantaneous introduction of the screening programme to all women in the invitation schedule would have led to a much sharper peak in the demand for radiotherapy.

screening history

Until not so long ago there was usually no need for an evaluation study of breast cancer screening in a Western European setting to take into account that there had previously been a significant amount of screening going on which influences the future screening effects which are to be evaluated. The first report on previous screening influencing a new screening programme is from Bouches du Rhone where it appeared that the programme has not led to a further improvement of stage distribution of breast cancer relative to the situation before screening in which already lots of opportunistic screening was taking place. (McCann et al. 1997)

Contrary to breast cancer screening, there is a much more extensive history of opportunistic screening in Western Europe concerning Pap smear taking for the prevention and early detection of cervical cancer and its mortality. A previous screening practice can influence the optimum of a future screening programme in several ways, making it necessary to take this history into account:

- The screening programme will in part be just a substitution of the existing practice, therefore leading to less extra effects than when the programme would be applied to a totally unscreened population. Reduction of incidence and mortality observed in the population due to the introduction of the programme will be less than the overall effect of screening. Also the effects on demand for health care facilities will be less salient.
- The age distribution of the screening history may have been far from optimal. For instance in cervical cancer screening we see that opportunistic screening tends to concentrate in younger ages where most pre-invasive lesions are detected instead of the middle and older ages in which most preventable cancer incidence and mortality occur. This may lead to the conclusion that, particularly at the start of the new programme, an extra effort needs to be made for sufficient screening of middle aged and older women.

screening attendance

In breast cancer screening, usually the cost of inviting people to a screening programme is low relative to the cost of the screening procedure as a whole, therefore the magnitude of cost and effects

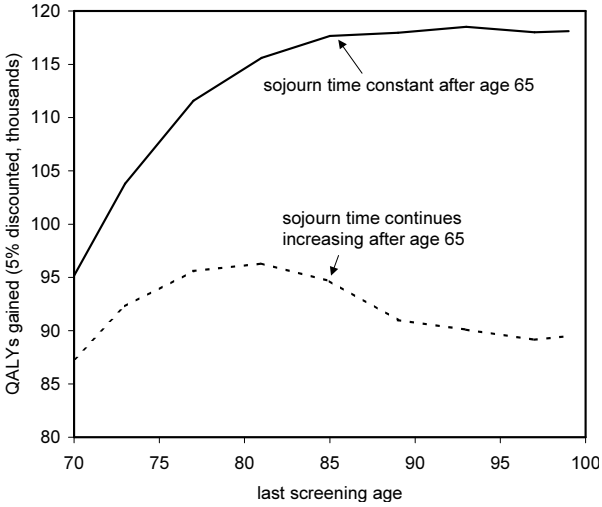
will be proportionally influenced by attendance, thus the cost-effectiveness ratio hardly depends on attendance. This will be very different for screening for colorectal cancer by a simple faecal occult blood test for which the cost of performing a screening test is not very different from sending out an invitation. There is a tendency that healthy people and/or people with generally healthy behaviour are more likely to attend to a screening programme than others.
In the pilot project on breast cancer screening in Nijmegen (Otten et al. 1996; Verbeek et al. 1984), we have observed in their database (Verbeek and Straatman, pers. comm.) that among women of 68 to 74, there is a major difference in mortality from other causes than breast cancer between those women attending to screening and those who do not attend, particularly in a short period after the invitation for screening. Women of around age 70 who attend to screening have at that time a 2 years longer life expectancy than those who do not attend, even without taking into account that the screening may prevent breast cancer death. This means that screening can gain more life years in older women than what would be expected without such selection effect.
In cervical cancer screening another selection effect is observed: women attending to screening have a lot lower risk for cervical cancer and its precursors than others women, even if there would be no screening effect. That means that the impact on mortality and incidence of cervical cancer in the total population will be markedly less than expected without such selection effect. For example: If attending to screening leads to a 90% reduction of cervical cancer mortality, then in a population where 80% of the women attend without association with their risk, mortality will go down by 80% * 90% = 72%. But if there is a 10% stratum of women who do not attend screening with a risk 3 times higher than other women, then mortality will only go down by (90%/(3 * 10% + 90%)) * 80% * 90% = 60%.

clinical practice
Diagnostic procedures in women with breast cancer are not likely to be very different among screening contexts in Europe: a breast cancer is diagnosed by a biopsy after a number of procedures involving markedly less burden on the woman and costs. For cervical cancer screening the practice of coming to a conclusive diagnosis may be more variable, but will probably still not lead to major differences in cost-effectiveness of cervical screening.
Due to variability of specificity, differences can be a lot greater when it concerns the diagnostics in women without cancer. Low specificity leads to more diagnostic procedures induced by screening; the practice of diagnostics per false positive case can also lead to more or less invasive diagnostics. On the other hand: it is to be expected that screening will reduce diagnostics outside screening and this effect can also be very different among evaluation situations.
In cervical cancer screening the variation of diagnostic practice concentrates in the so called 'borderline lesions'. The threshold for which cytological abnormalities are regarded as needing further follow up and which one not, can vary strongly from one cytological laboratory to the other, as well as that this threshold may vary strongly over time. It is very difficult to assess the effects of such differences because new thresholds are not founded by randomised trials for their effects on cervical cancer incidence and it is not well possible to derive these effects from observational data.
Also therapy practice can vary among screening contexts and may have a different dependency on stage at detection. For instance in cervical cancer the hysterectomy rate in pre-invasive conditions may vary strongly between regions (van Ballegooijen et al. 1995).

survival
Survival depends on stage at detection (by screening or clinically) and the quality of therapy. It is considered that among situations in which cancer screening is considered, variation in survival due to quality of therapy in small is comparison with variation due to stage at detection (Schrijvers et al. 1995). If quality of therapy would lead to important variation in survival, the effect in general is analogous to what was explained concerning stage distribution: a more favourable survival leads to less potential effectiveness. However when treating is not optimal it should be considered that treatment of early cases may be particularly unfavourable in comparison with optimal treatment; in that case the improvement of prognosis due to early detection decreases, and thus also the effectiveness of screening.

costs
Relevant costs are not only cost of screening itself, but also of all relevant diagnostic and therapeutic procedures, also medical costs of the life-years saved can be considered (Drummond et al. 1997; Russell et al. 1996; Siegel et al. 1996; Weinstein et al. 1996).
What we have seen so far at the evaluation of different breast cancer screening situations is that the cost factor which is of greatest influence on cost-effectiveness is whether screening is performed by specialised screening units with an intensive use of the mammographic facilities or as part of a radiologist practice with a relatively low intensity of use of the facilities (van Ineveld et al. 1993).
The unit costs of relevant diagnostic and therapeutic procedures which may vary most strongly between

Figure 10.4 Expected number of quality adjusted life years gained from the Dutch national breast cancer screening programme by upper age limit of a two yearly screening schedule starting from 50 years old for a model where preclinical screendetectable sojourn time remains constant from age 65 and for a model where this sojourn time continues to increase



screening situations are those of diagnostics after a positive screening test and of advanced disease. The cost of diagnoses after positive screening tests strongly increases with worse specificity. The cost of advanced disease as influenced by screening is on the one hand proportional to the number of cancer deaths prevented, therefore also roughly proportional to effectiveness. On the other hand this cost of course depends on the amount of medical care a women with advanced disease receives. This cost per patient is shown to be largely dependant on the average number of hospital days per patient with advanced disease (de Koning et al. 1994; Richards et al. 1993; van Ballegooijen et al. 1992b).

balance of favourable and unfavourable health effects

After considering all previous influences on the effects of cancer screening and trying to quantify them as good as reasonably possible, there comes a time of making up the balance. Not any cancer screening effort is free of unfavourable effects, therefore the first question to be addressed by making up a balance should be: are the favourable effects outweighing the unfavourable effects? There may be occasions where the answer to this question is obvious, thus needing not much further consideration. However even in breast cancer screening where the positive effects are so well established, this question can become highly relevant when considering different policy alternatives.

An example is considering what is the best upper age boundary for the invitation schedule of a breast cancer screening programme. With increasing age of screening participants, the probability increases to die of other causes than breast cancer before the cancer that can be detected early at time of screening would lead to death or even would be diagnosed after symptoms developed. Diagnosing cancer due to screening that would not be diagnosed without screening is a major negative effect of screening. The strong increase of such negative effects with age of screening cause that at some age, the balance of positive and negative effects as expressed by quality adjusted life years (QALYs) gained by screening becomes so unfavourable that further extension of a screening programme to higher ages does not further increase QALYs gained. This effect is a stronger with longer sojourn times of preclinical cancer. (Boer et al. 1995a) Figure 10.4 shows that if sojourn time does not further increase after age 65, the upper age limit at which the unfavourable effects are expected to outweigh the favourable effect is high, while if this sojourn time continues to increase, already around age 80 screening is not expected to gain any more QALYs.

For cervical cancer we don't know any good attempt to estimate the balance of positive and negative health effects of screening, that's why we have tried to extrapolate our findings for breast cancer screening to cervical cancer screening. Though this is merely a tentative estimate, it shows that the weight of negative effects of this screening can be quite substantial relative to its positive effects. (van Ballegooijen 1998b)

time preference

In balancing favourable and unfavourable health effects as well as effectiveness and costs, for a screening programme, time preference plays an important role. It is customary in cost-effectiveness analysis to express time preference as a yearly discount rate for both health effects and costs, though it has been argued that others models may be better. (Cairns and van der Pol 1997) Because the costs of cancer screening need to be made long before the main targeted effects (prevention of cancer death and possibly incidence) occur, the cost-effectiveness ratio is very sensitive for the applied discount rate.

Table 10.3 shows an arbitrary example of cost-effectiveness ratios of cervical cancer screening

table 10.3 Cost-effectiveness ratios of cervical cancer screening by applied yearly discount rate in a Dutch situation with an invitation policy of seven 5-yearly invitations from ages 25 through 55

yearly discount rate	cost effectiveness ratio Dfl/life year gained
0%	6,248
3%	16,788
5%	29,024
8%	58,399

which makes clear how strongly this ratio depends on discount rate.
Another problem with discounting arises when evaluation of screening concerns a so called cohort model as opposed to considering a full dynamic population. In such cohort approach it is assumed that all potential participants to a screening programme will be offered the complete invitation schedule and no-one will enter the programme at a later age than the first age of the screening schedule. Applying a discount rate then works out as age preference instead of time preference. This means that screening performed at older ages gets too low weight in the cost-effectiveness ratio as compared to screening at younger ages. For decision making (within the programme to be evaluated) concerning which particular ages are to by invited and which ones not, this is not a problem, because the marginal cost-effectiveness ratio of adding a screening age to the schedule will be correct, but in comparisons with other programmes it biases in favour of programmes which derived their cost-effectiveness from younger ages.

conclusion

When making prognoses for a future or starting screening programme, one should obviously start with interpreting the primary results of trials and observational studies. But for extrapolating the results of trials to other screening programmes and other situations, there is need for more data on background variables such as incidence and stage distribution in the situation without screening. Decisions on whether to start a cancer screening programme and on which screening policy is preferable, should not only depend on the finding that trials show a significant mortality reduction, but should also take several background variables into account concerning epidemiology in the situation without screening, demography, expected screening quality, clinical practice as well as the costs involved.

11. EXTRA INCIDENCE CAUSED BY MAMMOGRAPHIC SCREENING

The issue of extra incidence caused by mammographic screening plays an important role in the decision making of implementation of a population based screening programme. At the 'Seventh European Conference on Clinical Oncology and Cancer Nursing' of November 1993, Dr. Knut Aspegren reported that the Malmö breast cancer screening trial shows 25% extra incidence in the study group as compared to the control group when cumulated over 10 years of follow up from randomization.

We have used a breast cancer screening model which accounts for the important characteristics of the Malmö trial, such as age distribution, screening interval, and attendance.(de Koning et al. 1991; van Oortmarssen et al. 1990b) At 10 years of follow up the model predicts an increase of incidence of 31% of the screened group as compared to the control group. That is even higher than observed. However, when we define 'real' extra incidence as cancers detected in a screening programme which would not have been detected during life time without the programme, the same model predicts that the Dutch mass screening programme (inviting women every 2 years from age 51 to 69 and 70% attendance) causes less than 2% 'real' extra incidence in the total population.

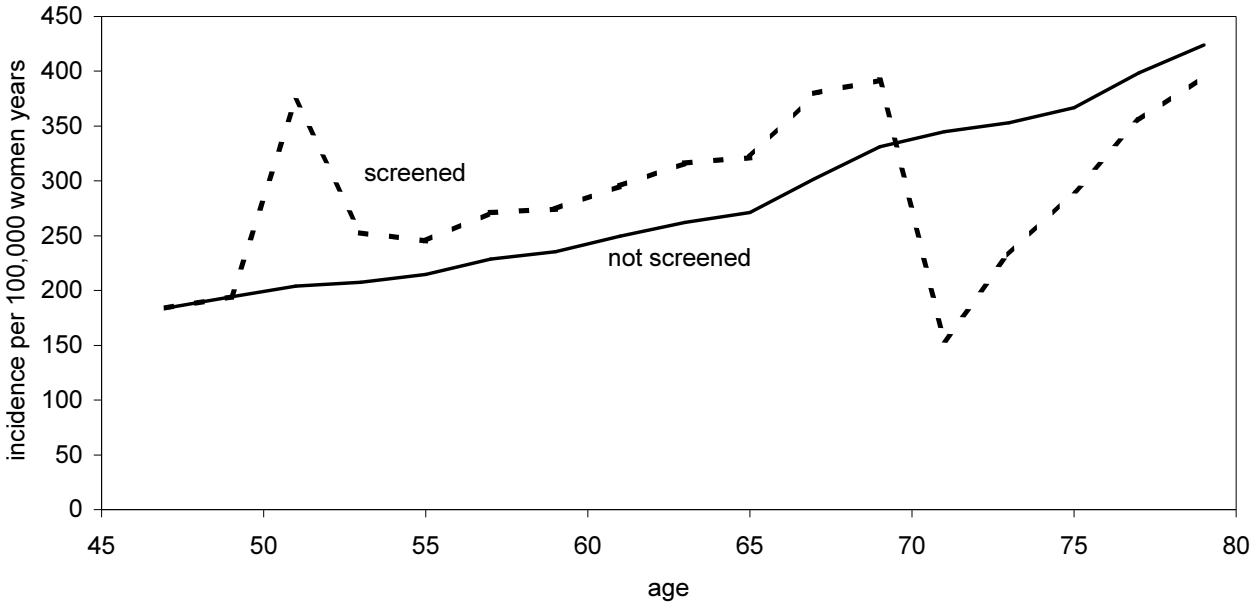
The figure shows the expected incidence of women who take part in the Dutch screening programme in comparison with women who do not take part in the programme. The first screening (at age 51) results in a sharp increase of incidence because of detection of the prevalence pool of breast cancers. At subsequent screenings there is still a marked increase of incidence because cancers are detected earlier by screening and incidence increases strongly with age. We estimate the average lead time in the Dutch programme to be 4.8 years on average and increasing with age.

However when the screening programme stops, incidence in screened women is expected to drop to a much lower level than in unscreened women. If the age at the last screening is sufficiently low, the decrease of incidence after the programme almost completely compensates for the increase of incidence during the programme. Thus the large observed extra incidence in a trial is mainly explained by early detection which is necessary for achieving a reduction of breast cancer mortality, and only a small proportion of the extra incidence would not have been detected without screening.

additional discussion

The letter to the editor of this chapter explains that a higher incidence in the screened arm of a trial relative to the control arm may very well be a very temporary effect while on the long run there is hardly any extra incidence due to breast cancer screening. The letter concentrates on how alarming the difference of incidence in the two trial arms is after limited follow up, but does not extensively go into estimating how large extra incidence caused by breast cancer screening is on the longer run. For the model that was used for this letter, we assumed

Figure 11.1 Expected breast cancer incidence in 2-year age categories



exponential distributions of the dwelling times of the preclinical disease states. This assumption has shown to be in agreement with observations in screening studies that measure detection rates and interval cancer incidence. However the amount of (long term) extra incidence is largely determined by the tail of the dwelling time distribution. The shape of the sojourn time distribution would appear as the difference between the interval cancer incidence by time since last screening and the expected incidence in a situation without screening. It is not possible to measure this difference with any satisfactory precision for long follow up after last screening. Therefore one may wonder if the model estimate for extra incidence has any precision. If one would consider dwelling time distributions to be Weibull with shape parameter 0.5 and the same average duration, then the model expectation for extra incidence caused by the current Dutch screening programme is 5.6% as compared to 5.2% expected extra incidence when assuming exponential distributions. That difference is so small because in order to have the same average duration, increasing the number of very slow growing cancers needs to be compensated by increasing the number of very fast growing cancers and they lead to very little extra incidence.

12. STAGE DISTRIBUTION AT FIRST AND REPEAT EXAMINATIONS IN BREAST CANCER SCREENING

abstract

objectives

Contrary to expected, stage distribution at repeat screenings for breast cancer is often not more favourable than stage distribution at first screenings. This article investigates observed stage distributions at first and repeat screenings. The observed outcomes are compared with expected values based on simulation modelling, varying the assumptions on natural history of the disease.

methods

An overview is made of observed data on stage distribution at first and repeat screenings and the difference between those distributions is summarised in a Gini coefficient. Four possible explanations for the observations are considered, two of these are worked out as Miscan simulation models of which the outcomes are compared with observations.

results

Half of the reported stage distributions at repeat screening are not more favourable than at first screenings and in the ones that are more favourable, the difference is relatively small. By assuming no correlation between the sojourn times of preclinical breast cancer in consecutive tumour size categories as well as a strong influence of latent cancers in a Miscan model, it is not possible to reproduce the observed outcomes.

conclusions

The two modelled explanations are not sufficient. Decreasing sensitivity seems an unlikely explanation for the discrepancy in many screening programs. False reassurance, as a remaining possibility to explain observations, should be seriously considered and investigated.

introduction

The performance of cancer screening programs can be measured by the detection rates (= numbers of breast cancers detected per 1000 women screened) and stage distributions in first and subsequent screening rounds and by the incidence and stage distribution of interval cancers in relation to a situation without screening. The stage distribution of cancers detected in the first screening round, as a consequence of earlier detection, is expected to be more favourable than that of clinically diagnosed cancers. Because most prevalent cancers are removed from the screened population at the first round, subsequent rounds will only detect cancers that have developed during the screening interval. If the screening interval is sufficiently short, detection rates in subsequent rounds will be much lower than in the first round, and the stage distribution is expected to be more favourable than in the first round. In breast cancer screening programs, detection rates in subsequent rounds have indeed been found to be much lower than in the first round, but the expected concomitant improvement in stage distribution is often small or even absent.

This article investigates the stage distribution of breast cancers detected at repeat screenings compared to first screenings as reported in the literature, using a simple summary measure for a difference in stage distribution. The observed outcomes are compared with expected values based on simulation modelling. Four possible explanations for the difference between observed and expected values are suggested, each of which is considered with respect to possibilities of further investigation.

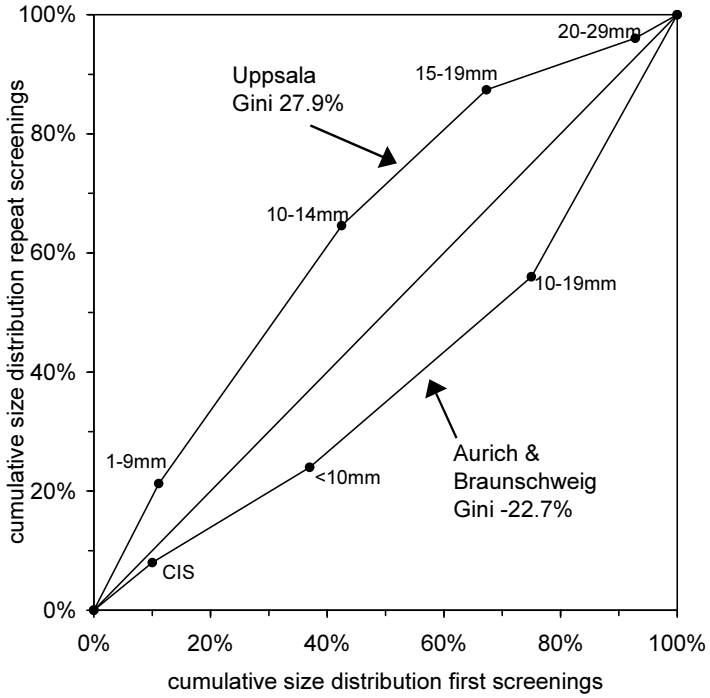
methods

Observed data were derived from publications on several breast cancer screening projects (Anderson et al. 1991; Andersson 1984; Boer et al. 1998; de Koning et al. 1995b; Duffy et al. 1991; Fagerberg et al. 1985; Frankel et al. 1995; Frischbier et al. 1994; Frisell et al. 1989; Frisell et al. 1986; Hakama et al. 1995; Libstug et al. 1998; Paci et al. 1995; Roberts et al. 1990; Tabár et al. 1984; Tabár et al. 1989; Tabár et al. 1992; Thurfjell and Lindgren 1994; van den Akker-van Marle et al. 1997), summarised in table 12.1. Some reports split up screen-detected cancers by first and repeat screenings, others by first and subsequent rounds where a subsequent round contains some first screenings of women who were not screened at the first round of invitations. Different categorisation of size distributions complicates comparison, therefore we summarised the difference in size distribution between first and repeat screening by a single value. The value used is the Gini coefficient, originally intended to measure inequality of income in a population (Ekelund and Tollison 1986). The Gini coefficient G is given by $G = \sum_{i=1}^{N-1} (p_i + p_{i+1})(Q_i - P_i)$, where p_i and q_i denote the proportion of cancers in tumour size category i for first respectively subsequent screenings; and $P_i = \sum_{j=1}^{i} p_j$ and $Q_i = \sum_{j=1}^{i} q_j$ are the corresponding cumulative proportions, N is the number of size categories.`

Figure 12.1 illustrates the calculation of this coefficient for two screening projects. The x-axis sets off the cumulative distribution of tumour sizes at first screenings of cancers with known size and the y-axis does the same for subsequent screenings. The more favourable the size distribution of repeat screenings is relative to first screening, the higher the graph will

be in the figure. The Gini coefficient is the area between the graph and the diagonal multiplied by 2, the area below the diagonal is counted as negative. When both size distributions are equal, the Gini coefficient is 0. Increasingly more favourable size distributions of repeat screenings relative to first screenings are reflected in higher Gini coefficients.
Figure 12.1 shows the diagonal which represents equal stage distribution and two extreme cases of a difference between size distribution at first and repeat screenings. The 95% confidence intervals for the Gini coefficients have been estimated by the bootstrap method (Efron and Tibshirani 1993). This method ignores any dependency of stage distribution at first and repeat screenings, but we expect that the influence of that on the variability estimate is negligible.

Figure 12.1 Illustration the Gini coefficient for two screening projects: cumulative distribution of tumour sizes at first screenings versus subsequent screenings



The computer simulation package Miscan for breast cancer screening evaluation is used for generating expected values for stage distribution at first and repeat screenings (de Koning et al. 1991; van Oortmarssen et al. 1990b). Miscan incorporates the natural history of disease, the epidemiology, the design of the screening program and performance of screening (Habbema et al. 1985). The natural history of breast cancer is modelled as a progression through a number of states. A life history starts with 'no breast cancer' before the onset of preclinical screendetectable disease. There is one pre-invasive state dCIS (Ductal Carcinoma In Situ) and 4 invasive states in the model, according to T-status of TNM classification (T1a, T1b, T1c, T2+).
In the basic model used in this paper, the duration in the different states follows an exponential distribution and durations of subsequent disease states are 100% correlated so that the duration of the total preclinical screendetectable period is also approximately exponentially distributed. The transition to the clinically diagnosed states (with the same subdivision) are governed by the age-specific incidence rates and the stage distribution in a situation without screening. In the case of early detection women will enter the screen-detected states (again with the same subdivision). The assumptions on the mean duration by age of the screen-detectable pre-clinical states of breast cancer and the sensitivity of screening have been validated using all data from the Dutch screening projects in Nijmegen and Utrecht (van Oortmarssen et al. 1990a). The mean duration of the preclinical screendetectable period increases with age from 1.8 years at age 35 to 6.2 years at age 70. At age 60 the mean durations of the different preclinical screendetectable disease states are 5.2 years for dCIS, 0.14 years for T1a, 0.72 years for T1b, 1.49 years for T1c and 1.12 years for T2+. The sensitivity of a mammographic screening assumed for the different disease states are 40% for dCIS, 65% for T1A, 80% for T1B, 90% for T1c and 95% for T2+.
Three variants of this basic model are considered in this paper.
Variant I assumes that there is no correlation between durations of consecutive disease states of a cancer. In order to maintain the same variance of the total preclinical period as in the basic model, model variant I assumes also that the duration in one disease state has a greater variance than in the basic model: a weibull distribution with shape 0.6. This means that a relatively large proportion of cancers have a very short dwelling time in a particular disease state, for instance: In an exponential distribution 5% of the cancers have a dwelling time of less than 5% of the average dwelling time, while a weibull distribution with shape 0.6 this is not 5% but 15%.
Variant II assumes that tumours that will be diagnosed in a situation without screening in a later stage, have a faster growth rate than cancers which are clinically diagnosed in an earlier stage. In model variant II the durations in subsequent disease states are chosen such that cancers which will be diagnosed as T2+ have on average 2 times shorter durations in their preceding disease states than those that will be diagnosed as T1c, while cancers diagnosed as T1c on their turn grow twice as fast as those diagnosed as

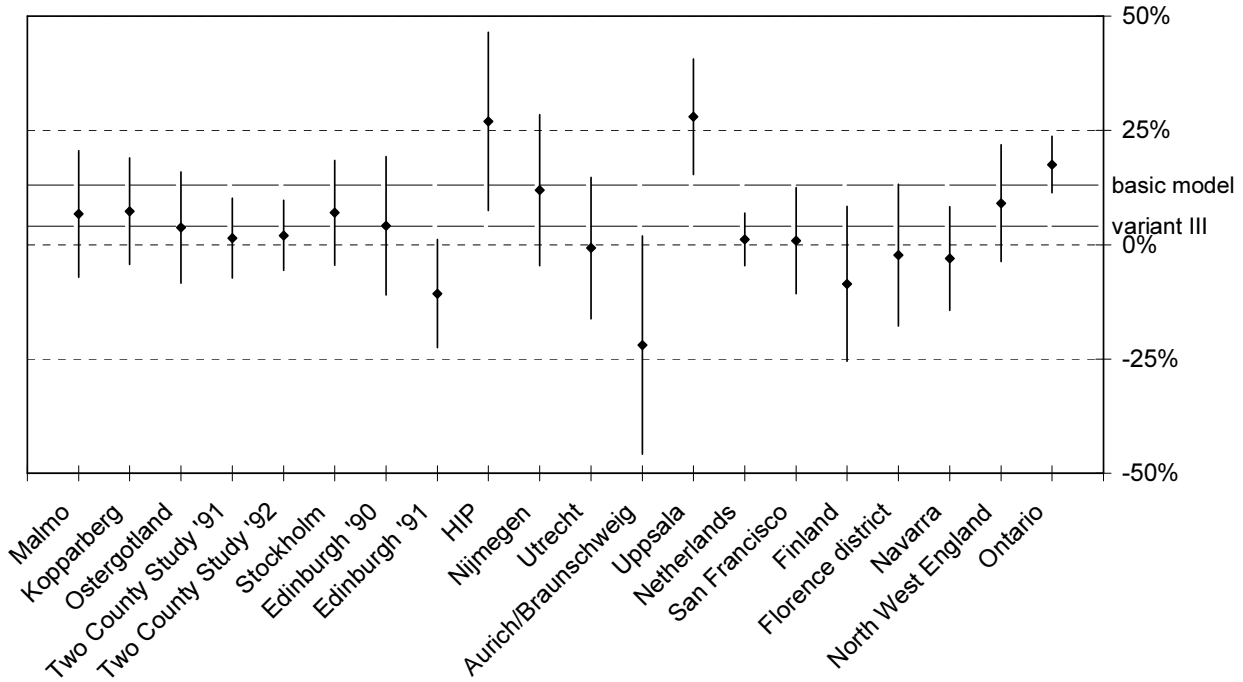
table 12.1 Overview of observed stage distribution by tumour size or UICC stage at first and repeat screenings

age range/ round	screening interval	number of cancers	stage distribution by size or UICC stage				
Malmo trial (Andersson 1984)							
45-69			Non-Inv.	Inv. <=10mm	Inv. >10mm		
round 1		118	16%	43%	41%		
round 2	22 mnths	58	21%	48%	31%		
round 3	20 mnths	46	15%	46%	39%		
Two County Study, Kopparberg (Tabár et al. 1984)							
>=40			Tis	<=5mm	<=10mm	<=20mm	>20mm
first round		275	7.6%	3.6%	28.4%	37.5%	22.9%
second round	24/33 mnths	109	10.1%	8.2%	20.2%	47.7%	13.8%
Two County Study, Ostergotland (Fagerberg et al. 1985)							
40-74			Tis	<=10mm	<=20mm	21-50mm	>50mm
1st screening		226	10.2%	38.5%	34.9%	14.2%	2.2%
2nd screening	24/33 mnths	106	11.3%	36.8%	42.5%	8.5%	0.9%
Two County Study (Duffy et al. 1991)							
40-69			1-9mm	10-14mm	15-19mm	20-29mm	30+mm
1st screening		284	26.1%	29.6%	19.7%	14.1%	10.5%
2nd screening	24/33 mnths	375	27.2%	26.4%	24.0%	16.5%	5.9%
Two County Study (Tabár et al. 1992)							
40-74			1-9mm	10-14mm	15-19mm	20-29mm	30+mm
first screening		382	26%	29%	19%	15%	10%
later screenings	24/33 mnths	424	27%	27%	24%	17%	6%
Stockholm trial (Frisell et al. 1989)							
40-64			Tis+I	II+			
first round		124	70%	30%			
second round	2.3 yrs	92	77%	23%			
Edinburgh trial (Roberts et al. 1990)							
45-64			<=10mm	11-20mm	21-50mm		
prevalence scr.		73	25%	47%	29%		
incidence scr.	1 yr (2yrs mammogr)	122	29%	44%	27%		
Edinburgh trial (Anderson et al. 1991)							
45-64			pTis	pT1	pT2	pT3	
prevalence scr.		123	19.5%	56.1%	20.3%	4.1%	
incidence scr.	1 yr (2yrs mammogr)	165	12.1%	58.2%	21.2%	8.5%	
HIP trial (van Oortmarssen et al. 1990a)							
40-65			<=20mm	>20mm			
initial screen		42	40%	60%			
repeat rounds	13 mnths	51	67%	33%			
Nijmegen pilot project (van Oortmarssen et al. 1990b)							
35-64			dCIS	<=9mm	10-19mm	>=20mm	
round 1		74	12%	16%	32%	39%	
round 2-4	2 yrs	120	8%	25%	43%	24%	

table 12.1 (cont.) Overview of observed stage distribution by tumour size or UICC stage at first and repeat screenings

Utrecht pilot project (van Oortmarssen et al 1990b)								
50-64			dCIS	<=9mm	10-19mm	>=20mm		
round 1		108	14%	25%	40%	21%		
round 2-4	1, 1.5, 2, 4 yrs	81	6%	32%	46%	16%		
Aurich and Braunschweig pilot projects (Frischbier et al. 1994)								
35-84			CIS	<10mm	10-19mm	20+mm		
1st screening		109	10%	27%	38%	24%		
2nd screening	1.1 yrs	25	8%	16%	32%	44%		
Uppsala screening program (Thurfjell et al. 1994)								
40-69			1-9mm	10-14mm	15-19mm	20-29mm	30-49mm	50+mm
1st round		153	11.1%	31.4%	24.8%	25.5%	4.6%	2.6%
2nd round	unknown	127	21.3%	43.3%	22.8%	8.7%	3.1%	0.8%
Dutch screening program (de Koning et al. 1995b)								
50-69			dCIS	T1a	T1b	T1c	T2	T3,4
1st screening		1754	14.7%	6.3%	21.1%	37.9%	17.9%	2.1%
2nd screening	2 yrs	449	14.4%	5.6%	23.3%	37.8%	17.8%	1.1%
San Francisco screening practice (Frankel et al. 1995)								
age distribution unknown			0	I	II	III	IV	
initial screening		222	30.2%	46.8%	21.2%	0.9%	0.9%	
subs. screening	n.a.	111	26.1%	56.8%	15.3%	1.8%	0.0%	
Finnish screening program (Hakama et al. 1995)								
50-64			Tis	T1	T2-3			
1st round		135	14.1%	64.4%	21.5%			
2nd round	2 yrs	39	7.7%	66.7%	25.6%			
Florence district screening program (Paci et al. 1995)								
40-69			Tis	T1	T2	T2+		
1st screening		77	9.1%	50.6%	31.2%	9.1%		
2nd screening	2.5 yrs	126	6.3%	50.1%	39.7%	3.9%		
Navarra screening program (van den Akker-van Marle et al. 1997)								
50-64			CIS	T1a	T1b	T1c	T2+	
1st screening		286	16.6%	4.7%	22.4%	33.2%	23.1%	
2nd screening	2 yrs	136	11.9%	7.5%	23.9%	31.3%	25.4%	
North West England screening program (Boer et al. 1998)								
50-64			dCIS	micro-invasive	other T1a	T1b	T1c	T2+
1st screening		392	15.1%	2.6%	6.6%	30.6%	30.4%	14.8%
2nd screening	3 yrs	92	20.7%	4.3%	7.6%	27.2%	28.3%	12.0%
Ontario screening program (Libstug et al. 1998)								
50+			in situ	micro-invasive	<10 mm	10-14 mm	15-19 mm	20-49 mm	50+ mm
initial screen		1189	13.3%	2.6%	17.7%	23.3%	15.8%	25.6%	1.8%
rescreens	2 yrs	336	18.8%	6.0%	20.8%	22.0%	17.6%	14.6%	0.3%

Figure 12.2 Gini coefficients for all screening projects, bars indicating the 95% confidence interval, horizontal dashed lines indicating model predicted Gini coefficients



T1a or T1b. The prevalence of each disease state is chosen equal to the basic model. That means that of those cancers that are clinically diagnosed as at most T1b, the average duration in T1a at age 60 is 0.34 years and in T1B 1.68 years; of those that are clinically diagnosed as T1c, that duration of T1a is 0.17 years, in T1b 0.84 years and in T1c 2.18 years; and those that are clinically diagnosed as T2+, that duration in T1a is 0.09 years, in T1b 0.42 years, in T1c 1.09 years while the duration in T2+ of course does not change (1.12 years) because they are all clinically diagnosed as T2+.

Variant III combines the adjustments to the basic model of both variant I and II by assuming no correlation between consecutive disease states compensated by a greater variance per disease state than in the basic model: a weibull distribution with shape 0.7 (slightly larger than in variant I) and the same average durations of disease states as in variant II.

The differences between these model variants do no affect stage distribution in a situation without screening.

results

Table 12.1 shows the overview of available data on stage distribution by tumour size at first and repeat screenings. In the first column of table 12.1 is indicated whether the data concern actual first versus repeat screenings or first versus repeat screening rounds where the repeat screening rounds can contain some first screenings. Figure 12.2 shows the summary of these data by the Gini coefficients from which it becomes clear that in by far most screening

table 12.2 Expected values for the stage distribution at first and repeat screening and for the Gini coefficients, for basic model and the three model variants

		Dcis	T1a	T1b	T1c	T2+	Gini c.
stage distribution in situation without screening		4.1%	1.5%	6.9%	33.2%	54.4%	
basic model	1st screening	16.4%	3.6%	21.0%	40.5%	18.5%	13.0%
	2nd screening	14.4%	6.0%	33.1%	37.6%	8.9%	
variant I	1st screening	15.6%	3.7%	20.8%	40.3%	19.5%	8.0%
	2nd screening	13.7%	6.1%	28.5%	37.6%	14.1%	
variant II	1st screening	16.5%	3.3%	18.9%	41.5%	19.8%	9.0%
	2nd screening	13.6%	5.5%	28.5%	39.7%	12.7%	
variant III	1st screening	16.4%	3.7%	18.7%	40.8%	20.4%	4.0%
	2nd screening	12.8%	5.3%	25.9%	39.1%	17.0%	

projects, stage distribution at repeat screenings is not much better, often even worse, than at first screenings.
The Two County Study shows a clearly more favourable tumour size distribution at repeat screenings than at first screenings (Duffy et al. 1991; Fagerberg et al. 1985; Tabár et al. 1984; Tabár et al. 1992). The trials in Malmo (Andersson 1984) and Stockholm (Frisell et al. 1989) are reporting similar differences, which are not so large. The different publications from the Edinburgh trial do not show unequivocally whether stage distribution is more favourable at first than at repeat screenings (Anderson et al. 1991; Roberts et al. 1990). Also screening projects in Italy (Paci et al. 1995), Germany (Frischbier et al. 1994), Finland (Hakama et al. 1995), California (Frankel et al. 1995) and in Spain (van den Akker-van Marle et al. 1997) do not show a more favourable stage distribution at repeat screening than at first screenings.
Only from the HIP trial (van Oortmarssen et al. 1990a) and from projects Uppsala (Thurfjell and Lindgren 1994), the U.K. (Boer et al. 1998) and Ontario (Libstug et al. 1998) there are reports of a substantially better stage distribution at repeat screenings in comparison to first screenings. In Uppsala there was an important increase in screening performance from the second screening round onward, which is illustrated by the fact that the detection rates in the first and second round of screening are equal. This explains why relatively few small tumours were found at first screenings.
There is no apparent correlation between screening interval and difference between stage distribution at first and repeat screenings, while one would expect that a longer screening interval would lead to a worse stage distribution at repeat screenings.
Summarising: of the 16 screening projects from which data have been studied here, 8 show a better stage distribution by tumour size at repeat screenings, 7 show a worse distribution and from 1 trial, one report shows a better and another shows a worse distribution.
Table 12.2 gives the expectations from the Miscan model for the Dutch national breast cancer screening program. The basic model clearly expects a much more favourable stage distribution at repeat screening compared to first screenings than is observed. The model predicts a Gini coefficient of 13.0% while 1.2% is observed with a 95% confidence interval of -4.6% to +7.0%. Model expectations for other screening projects would be very similar to the one for the Dutch program. Differences in model expectations would be due to differences in screening interval and ages of screened women.
Model variant I (without correlation between durations of subsequent disease states) shows a substantially less favourable expected stage distribution at repeat screening than the basic model while the distribution at first screening is very similar. The Gini coefficient is 8.0%.
Model variant II (where cancers which are diagnosed in later stages have a higher growth rate) shows similar stage distributions as variant I. The Gini coefficient is 9.0%.
The combined effects of adjustments from model variants I and II as described by model variant III further worsens the stage distribution at repeat screening leading to a Gini coefficient of 4.0%.
The Gini coefficient from variant III and the basic variant of the model for the Dutch program are signified in figure 12.2 for easy comparison of observed and model outcomes.

discussion
The observed overall similarity of stage distribution at first screenings and repeat screenings is different from the general expectation. We have tried to quantify this expectation by using the model that was validated against all data from the Nijmegen and Utrecht screening projects. Most observed Gini coefficients fall far below the expected 13%. We can conceive four possible explanations for this discrepancy:

little correlation between dwelling times of consecutive disease states
The basic model assumes a large variance in rate of development of screendetectable preclinical cancer: the time spent between becoming screendetectable and clinical diagnosis follows an exponential distribution. However, given a certain dwelling time in preclinical cancer for an individual, the proportion of time spent in each of the different preclinical disease states is fixed. This is clearly a simplification. When still assuming exponential distribution of the total screendetectable period, but no correlation between subsequent preclinical disease states, the time spent in one preclinical disease state will therefore have a more than exponential variance. Such higher variance in dwelling times in earlier disease states will lead to a larger proportion of cancers having grown into a larger state during the interval between screenings, thus to a less favourable stage distribution at repeat screenings.
Therefore, instead of the 100% correlation between durations in preclinical disease states as in the basic model, model variant I assumes the other extreme: no correlation. This model leads to an expected Gini coefficient of 8% which is still higher than most observed Gini coefficients.
The correlation between dwelling times as discussed in the previous paragraphs practically cannot be observed because it requires more than two observations of the disease state during the preclinical development of the cancer. Further

research therefore is necessarily limited to theoretical approaches. Perhaps this theoretical approach in future will lead to testable hypotheses.

latent cancers

Another explanation is that of latent cancers which are much more prevalent at first screenings than at repeat screenings. Latent cancers may be cancers that can be detected by screening, but which would never be clinically diagnosed in a situation without screening, or they are cancers with a very long preclinical screendetectable period. The first of these two groups would result in a non-transient increase of incidence due to screening. Such increase is not observed (therefore is at most small) and assuming an exponential distribution of the preclinical screendetectable period as in the basic model accurately predicts observed temporary changes in incidence (Boer et al. 1994).

The effect of latent cancers which is already in the basic model can be further increased by assuming a correlation between duration of preclinical disease and stage at clinical diagnosis.

If fast growing cancers are, on average, clinically diagnosed in a later stage than slow growing tumours, then the stage distribution at repeat screening is expected to be worse than in the basic model.

We do not know of any attempt to measure a correlation between growth rate and stage at diagnosis outside screening. However we think that the assumed ratio of growth rates of 1:1:2:4 for cancers that will be diagnosed as T1a, T1b, T1c or T2+ as in model variant II represents a correlation at the high side of a plausible range. The expected Gini coefficient of 9% is again higher than most observed Gini coefficients. Assuming a higher correlation would imply extremely short dwelling times in disease stages preceding the stage in which the cancer is diagnosed without screening, thus a low probability to be detected at screening in an earlier stage than if there would be no screening. This does not agree with the fact that the stage distribution of screen-detected cancers is generally much better than that of clinically diagnosed cancers.

It is possible to investigate this correlation in studies like those by Peer et al (Peer et al. 1993) or Brekelmans et al (Brekelmans et al. 1996). A correlation between growth rate and tumour size at diagnosis can be determined from the size at diagnosis outside screening and the apparent size of the tumour as it appears on one previous mammogram. Obviously clinically diagnosed cancers that can be seen on a previous mammogram will form a selected sample, but this does not necessarily invalidate the test for correlation between growth rate and stage at diagnosis.

Even the combination of the two previously described assumed extreme model variants into model variant III, with a Gini coefficient of 4%, still leads to an expected stage distribution at repeat screenings which is more favourable than at first screenings.

false reassurance

A negative screening result may induce false reassurance in a woman who may postpone a visit to the doctor for a lump she may feel in her breast and the doctor who may postpone further diagnostics. Such patient's or doctor's delay may lead to a later diagnosis of interval cancers and to detection of cancers at repeat screenings which otherwise would have been diagnosed earlier as an interval cancer. For this article, only the last possibility is relevant because it influences stage distribution at repeat screenings.

The stage distribution of interval cancers is less favourable than that of screen-detected cancers. A delay in the detection of such cancers up to the moment of a next screening will thus worsen the stage distribution in subsequent screening rounds. The current Miscan model cannot model false reassurance.

Delay in diagnosis of breast cancer has been studied earlier (Bedell et al. 1995; Caplan and Helzlsouer 1992; Caplan et al. 1996) particularly also its unfavourable effect on prognosis (Elwood and Moorehead 1980; Neave et al. 1990; Porta et al. 1991), the influence of a negative preceding screening on this delay has been mentioned already long ago (Burns 1978) but still urgently needs to be investigated empirically.

decreasing sensitivity

The Dutch population screening program performs double view mammography at first screenings and single view at repeat screenings. Also it is alleged that there is a pressure on the program to increase the number of mammographies made per screening unit which may have led to a decreased quality of work since the start of the program. These two factors may result in a lower sensitivity in repeat screenings, particularly for smaller cancers, therefore to a less favourable stage distribution. However it is not very plausible that these effects have occurred in so many screening projects all over the world.

It is possible to investigate whether first and repeat mammograms have the same technical quality, are interpreted according to equal standards and if two view mammography in repeat screenings will give a better sensitivity in particular for smaller cancers. The latter has already been demonstrated for first screenings (Blanks et al. 1996). It is also possible to investigate a possible difference in sensitivity after first and repeated screenings by comparing the

incidence of interval cancers after first and repeat screenings.

conclusion

The stage distribution at repeat examinations in most breast cancer screening projects is less favourable than expected. Four possible explanations are presented in this paper. Two of those ('little correlation between dwelling times of consecutive disease states' and 'latent cancers') have been shown to be at most part of the explanation for the discrepancy between observed and expected distributions, not even when both are working together; only the 'latent cancers' can be empirically investigated. Decreasing sensitivity during a screening program seems an unlikely explanation for the discrepancy in all screening programs considered in this article. While the reader is urged to present other explanations, there appears to be reason for serious concern about a substantial effect of false reassurance.

13. CONCLUSION

evidence from different randomised trials

There exists a strong need for integration of the knowledge that has been acquired through different studies. The meta-analysis tries to provide this information and has been applied quite successfully. It is primarily being used to increase the precision of the main outcome measure of studies. Major problems with the method are the handling of differences between the studies and the fact that the primary outcome is not necessarily identical with the value of greatest interest for decision making. The joint analysis with Miscan as described in chapter 2 can be seen as an alternative method for the meta-analysis that can overcome some of the problems with joint analyses.

The meta-analysis can handle differences between studies by excluding studies that are not in agreement with a more or less narrow definition, thus limiting the meta-analysis to the most similar studies. This practice goes at the expense of the power of the analysis to increase precision. It also introduces the possibility of bias by assuming a definition of trials that are to be included that coincides with trials with an on average higher or lower outcome than the excluded trials. Besides that, the main outcome of a trial on cancer screening is typically a relative risk of dying from the disease in question in the screened/invited arm versus the control arm during the course of the trial. For decision making, reduction of mortality from the cancer and life years gained, both in the target population instead of the trial population, are more interesting values.

The Miscan model for breast cancer has been applied to analyse the outcomes of the Swedish breast screening trials while accounting for several characteristics of the different trials to estimate the model parameter 'improvement of prognosis' that represents the (tumour size dependent) probability to prevent breast cancer death due to detection by screening. This parameter is then used to estimate mortality reduction and life years gained in the target population for which screening is being considered. Thus, the Miscan approach tries to solve both the problem of joining data gathered under different circumstances and the problem of extrapolating trial results to a situation of decision making. We currently plan to do a similar joint analysis of trials for colorectal cancer screening.

A clear disadvantage of this Miscan approach is that relative to the method of meta-analysis, the joint analysis with Miscan clearly lacks statistical validity. The problem that precludes statistical validity of estimates by Miscan results from the amount of assumptions that are necessary to construct the model and of which the uncertainty is not known well enough. The first attempt to estimate uncertainty of a Miscan model did not resolve all problems. Due to the nature of the unresolved problems, it cannot be expected that they will be resolved fully acceptably in foreseeable future. (see also further in this chapter under 'evidence and uncertainty').

From the perspective of decision making there is the choice between extrapolating the statistically valid results from a meta-analysis to the situation concerning the decision to be made, either by applying a formal model or by a more informal method, and extrapolating from individual studies by a joint model analysis to the decision situation. In other words: statistical validity is inevitably lost due to the extrapolation, whether this extrapolation is preceded by a meta-analysis or whether the gathering of evidence is included into a joint model analysis. While the problem of loss of statistical validity remains, there is need for both the practice of meta-analysis in order to meet requirements concerning evidence based medicine, and for further assessment of the technology in question based on a joint model analysis.

case control studies into efficacy of screening

Observational studies can be used as alternative for randomised trials on efficacy of screening. For this purpose, the case control design has been adapted for screening evaluation. However, a widely recognised problem is the self-selection bias due to a likely risk difference between those who tend to undergo screening and those who do not. Originally the case control study is used for reasons of efficiency because it only samples a small fraction of the vast majority of individuals who do not have the disease. The adapted version for estimating screening efficacy tries to avoid the bias that arises from the fact that individuals who die from the disease have a period from diagnosis in which they are not screened. Chapter 3 shows that this bias is not sufficiently resolved by the case control methodology as it is being used. Further development of this methodology will probably be able to further diminish this healthy screenee bias, but it will not be able to prevent self-selection bias. This precludes the outcomes from case control studies on screening efficacy to form a sufficient evidence base for deciding to start a screening programme. After efficacy of screening has been demonstrated by randomised trials, there may be a role for case control studies to support additional technology assessment, such as the evaluation of screening in a small age range within the potential target population.

estimating net survival

Besides an estimate of the efficacy of screening, decision support requires several other parameters to be estimated that influence the effectiveness of a screening programme. Many of the relevant

parameters can only be estimated by observational studies. Probably the most important of these values is net survival from the disease in question. There is no gold standard for estimating net survival, therefore an evaluation of bias in estimating net survival is limited to comparing different, but all potentially biased, methods. Chapter 4 shows that the most heavily debated issues in estimating net survival, such as the quality of registration of cause of death, are of limited consequence, at least in the examples of colorectal cancer and prostate cancer. Other issues, such as whether to limit the analysis to first cancers in a patient, are at least as important. In principle all estimated survival values used should be evaluated for bias.

evidence and uncertainty

Sensitivity analyses have been applied in several part of this thesis. In sensitivity analysis, the effects of changing assumptions on the outcomes are studied, usually by changing one model parameter at a time and thus studying the effects of uncertainty of one parameter at a time. Sometimes the effects are studied of changing more than one parameter at the same time in order to study the joint effect of uncertainty arising from several parameters and their interactions.

In complicated models such as the ones used in this thesis, it is not feasible to fully explore all interactions between variations in parameters. It is more feasible to perform an uncertainty analysis. In such an analysis, a probability distribution is assumed for each of the parameters that are to be subjected to sensitivity analysis. The probability distribution of a parameter represents uncertainty concerning that parameter. For simpler models it is possible to derive the probability distribution of an outcome measure from the distributions on assumptions. For more complicated models, the probability distribution of an outcome measure can be accurately estimated by sampling from the probability distributions of model parameters and evaluating the model for each sample. The rigour of the uncertainty analysis gives rise to several questions concerning uncertainty in modelling for decision support that are also applicable to any other form of sensitivity analysis.

There is a general recognition that decision makers should be provided with a notion on how uncertain expected effects of a decision are, even if the decision maker is not readily interested in such information. However, it is not clear what should be the role of uncertainty in decision making. While on the one hand it can be argued that a decision should be based on just the expected effects, on the other hand there is the argument that uncertainty is to be avoided. If uncertainty is to be avoided, the question rises: what negative value is to be attributed to uncertainty? In other words: if there is a choice between a decision for something with a certain cost-effectiveness then how much more favourable should the expected value of an alternative with less certain cost-effectiveness be? Sometimes the desire to avoid uncertainty is described as avoiding risk. However, in decision support where results concern the balance between costs and effectiveness, the question rises what risk is to be avoided.

As long as it is not sufficiently clear what actual decision criterion will be used, it is neither clear what uncertainty should be presented, that of the cost-effectiveness of one policy, the marginal cost-effectiveness of one screening policy relative to a slightly less intensive policy, the policy being Pareto optimal relative to all other possible policies or the policy to be preferred at a particular threshold for (marginal) cost-effectiveness?

In uncertainty analysis, all uncertainty on assumptions is described in the same format of a probability distribution. These uncertainties can however be of quite different quality. For instance in screening evaluation, there is uncertainty due to limited numbers of observations, to recent developments that cannot be observed in a situation with widespread screening so that they can only be estimated with an uncertain model on screening influence, to future changes in the epidemiology of the disease, to actual screening behaviour (attendance, intervals, follow-up), and to outcomes of future negotiations on costs. It is unclear as to how far these different kinds of uncertainties should be treated as being equivalent.

balancing favourable and unfavourable effects at high ages

Chapter 8 shows that while in general the balance between favourable and unfavourable health effects of breast cancer screening is good, when screening is applied in higher age groups, the unfavourable health effects may easily outweigh the favourable health effects. We have not estimated this balance for screening for other cancers. If early detection of prostate cancer has a similar effect as that of breast cancer, then the balance of favourable and unfavourable effects probably changes at earlier age. That is because the longer sojourn time of invasive prostate cancer, particularly at higher ages, results in detecting more cancers that would not have been diagnosed without screening, and in more life years with cancer. Cervical cancer and colorectal cancer have a relatively short sojourn time of invasive cancer, and a relatively long sojourn time of precursors of cancer by which invasive cancer can be prevented. This complicates the situation too much for extrapolation of our findings on breast cancer screening.

effectiveness and circumstances

Chapter 10 shows that specific local circumstances can influence cost-effectiveness of screening programmes. However, several Miscan evaluations of breast cancer screening in different countries did not result in widely different outcomes as to the preferred screening policy. For instance Spain has a much lower breast cancer risk than the Netherlands, warranting less intensive screening. Miscan modelling was applied in two regions of Spain, but these regions had a much higher breast cancer risk than the average for Spain. Apparently regions with the highest breast cancer risk tend to be the ones wanting to start a breast cancer screening programme so that preferred policies among those regions are very similar.

Among regions with a substantial screening effort for cervical cancer, the differences between screening policies are quite large and they are by no means justified by differences in local circumstances on cost-effectiveness. One may wonder if this unjustifiable divergence of screening policies is due to a lack of firmness of the evidence for the efficacy of cervical cancer screening. Moreover, the main regional differences in cervical cancer screening concern the higher risk in developing countries like India and Brazil compared to that in western countries. Particularly in high-risk regions hardly any screening takes place.

future developments

The effects of cancer screening depend on a rather complicated process of development of the cancer in question as well as on several circumstances in which the screening takes place. A quantitative assessment of cancer screening is therefore only possible with the aid of an integrative model. Miscan is an example of such a model. The statistical validity of estimates from such models is as yet not very satisfactory. However, the formal description of all assumptions and the mechanism by which estimates from the model are derived, is superior over alternative methods (Isaacs and Fitzgerald 1999) because in theory it opens the possibility to discuss and criticise all aspects of the assessment. In practice, the formal description of model estimates is apparently too intricate for appropriate critique and thus they are rather light-heartledly either firmly accepted or equally firmly rejected, largely depending on whether the estimate is in agreement with the personal preference based on informal estimates that are not liable to detailed critique. This practice may improve because the number of research groups that is seriously working on this type of modelling is increasing, giving more possibilities for competitions and mutual criticism. There are also signs for a stronger interaction between model development and establishing empirical evidence.

Our research group is currently developing a new model for evaluating breast cancer screening that is better able to explain improvement of prognosis due to early detection and that can better separate the natural course of the cancer, the behaviour of the woman with the cancer in response to sign from the disease, and the effects of a screening programme. This model will provide a better tool to study the influence of changes in the earliness of diagnosis outside screening on the effectiveness of screening and the possible effects of delay of diagnosis due to a previous negative screening. In breast cancer screening there is also need for more detailed analysis of the randomised trials, particularly with respect to the effect of screening women under age 50. Such analysis should of course include the Canadian NBSS trial.

In cervical screening the Miscan model can be used to estimate the natural history parameters of HPV infections that may cause cervical cancer. Though the causal relationship between HPV infection and cervical cancer has been firmly established by now, it is still not clear whether some form of HPV screening can be more efficient than the current pap test.

The Miscan model for prostate cancer screening will continue to be used as an aid to interpret data from the ongoing trials and population trends, and will be used to help estimate the balance of favourable and unfavourable health effects of screening.

The Miscan model for colorectal cancer screening will be used to help design a trial in the Netherlands. This trial design is complicated by the choice between rather different screening tests that each would perhaps call for different screening intervals and screening ages. We are also intending a joint analysis of trials that have already been performed in other countries, giving an opportunity to further develop this method of analysis in another disease than breast cancer.

Recent developments in the possibilities to detect lung cancer and of infections that may cause stomach cancer, may lead to consider new screening programmes. The decision for such a new screening programme can be supported by for instance Miscan models.

Models such as Miscan will probably gain in importance for gaining rational control over complex decision situations that are inevitable where cancer screening is at stake and where the aim is not just to maximise the targeted effect, but to reach the best balance of favourable and unfavourable effect.

SUMMARY

There are several steps between empirical research and decision support for cancer screening. The steps of gathering evidence for efficacy of screening, evaluation of uncertainty on estimates for parameters and effects, striking of the balance between favourable and unfavourable effects, and the estimation of the influence of local circumstances under which screening is applied on the effects of cancer screening are advanced in this thesis on the basis of several examples. Most examples concern the application of the micro-simulation model Miscan that has been developed particularly for the evaluation of early detection of cancer.

gathering evidence

In 1993 an overview of the five Swedish randomised trials on breast cancer screening was published. At the time, these trials were considered to be the most important contribution to the evidence for efficacy of breast cancer screening. Chapter 2 describes a joint analysis of these trials that accounts for the variation in circumstances between the different trials. The analysis is aimed at estimating parameters of a breast cancer model that can be used for extrapolation of the evidence from these trials to other screening situations. This chapter also addresses the issue of the extent to which the observed breast cancer mortality reduction among women aged younger than 50 years at entry into the trial can be explained by screening that was applied at ages over 50 years. It is concluded that the improvement of prognosis due to early detection is larger than an earlier estimate and that the data published thus far possibly implied that the efficacy of screening women younger than 50 years is very small. The latter conclusion provoked the publication of data showing that the assumptions from which that possibility was derived, were not correct. More recently, the evidence for efficacy of breast cancer screening in women younger than 50 years has become much stronger. However, the question for cost-effectiveness of breast cancer screening in younger women still awaits a sufficient answer. The model based joint analysis as presented in chapter 2 can be regarded as complementary to the method of meta-analysis in which differences in circumstances among trials and extrapolation to other circumstances are problematic, but that has much better statistical validity.

Efficacy of screening is not always estimated from randomised trials but can also be estimated from observational studies. The most appropriate method for that is a case-control design with an adaptation to account for the fact that screening is not performed after diagnosis of the disease. Chapter 3 investigates bias in estimates of mortality reduction due to screening when a case-control design is used. It shows that apparently diminutive differences in matching of cases and controls, and in the counting of exposure to screening, can lead to strong bias. Even with the right matching and counting of exposure, an important amount of bias remains because the adaptation for exception from screening of prevalent cases is not quite accurate. Furthermore, there is possibility for bias due to association between participation with screening and risk for the disease. Therefore, case-control studies should only play a limited role in the evaluation over cancer screening.

Chapter 4 describes an investigation of an alternative method of estimating survival from a diagnosis of prostate cancer. This method is not based on a population of diagnosed cases, but on cases of death occurring during a limited period in time that are followed back to a precursing diagnosis of prostate cancer. If this method is applied to a stable situation, it produces accurate results, but if the situation changes over the course of the study period, for example due to ageing of the population or introduction of screening, then the estimate will be considerably biased.

Chapter 5 compares several more generally accepted methods of estimating survival. It shows that the most important differences between estimates for net survival, at least for colorectal cancer and prostate cancer, are due to the choice for either disease specific or relative survival (1.4% difference in survival percentage in colorectal cancer and 6% in prostate cancer) and the choice for either inclusion or exclusion of people with more than one primary cancer (about 4% difference). De differences due to other variants of estimation were even smaller. Nevertheless it is recommendable to evaluate a particular method of estimating survival for bias.

evidence and uncertainty

In the United Kingdom where women aged 50-64 are invited every three years for breast cancer screening, there appeared to be a relatively high incidence of interval cancers in the third year after screening. This raised the question whether the screening interval should be reduced from three to two years. Chapter 6 researches whether applying this higher screening frequency to the British mass screening programme for breast cancer is cost-effective. The estimated cost per life year gained of this intensification appeared to be not much higher than that of the existing programme. But from a perspective of cost-effectiveness, it is equally recommendable to raise the last age of invitation from 64 to 69 years.

Chapter 7 compares the cost-effectiveness of cervical cancer screening by cytological smears with screening by direct visual inspection of the cervix. This comparison is particularly interesting for developing countries because direct inspection is far

less costly and also does not require highly developed technological provisions. However this method of screening is less sensitive for abnormalities and result into a rather large percentage of false positives that require further diagnosis. When comparing equally effective screening programs, the total cost as well as requirement for technology is expected to be higher than with application of the pap smear. However, other methods of visual inspection with better test characteristics, for instance after treatment of the cervix with diluted acetic acid or with aid of a gynoscope, are possibly more cost-effective than the pap smear.

balancing favourable and unfavourable effects
Chapter 8 investigates the age at which further invitations for breast cancer screening result in more unfavourable than favourable health effects. Under very pessimistic assumptions this upper age is at 75 years. This result was an important reason for the extending the current Dutch national screening programme from upper age of invitation of 69 to 75 years. Under less pessimistic assumptions, this extension is also efficient from a cost-effectiveness perspective.
Chapter 9 considers whether the unfavourable effects of breast cancer screening at higher ages can be limited by a longer interval between the invitations for screening. When reducing the screening interval, most of the additionally detected cancers have a relatively short preclinical sojourn time. Early detection of such cancers can often prevent mortality from the disease, while it only rarely induces extra incidence. That's why unfavourable health effects increase less than favourable health effects when reducing the screening interval. The ratio of favourable and unfavourable health effects therefore is no reason to apply a longer screening interval at higher ages. The cost per life year gained are of course still higher for screening with higher frequency.

effectiveness and circumstances
Chapter 10 describes the importance of various circumstances in prospective evaluation of effects and costs of cancer screening. It describes the influence of epidemiological characteristics such as level of incidence, stage distribution and survival in the situation without screening, demography, the quality of the screening offered, the screening schedule, the screening that has already been applied before them moment of prospective evaluation, the attendance to screening, the clinical practice outside the screening programme, the costs of diagnostics and therapy, the balance of favourable and unfavourable effects, and the rate of time preference. Though the differences in circumstances of screening can be of great importance for the selection of a particular screening policy, they seem to have little association with differences in screening as applied in practice.

compelling questions
During the period of screening, the Malmö trial on breast cancer screening showed around 35% more breast cancers in the screening arm than in the control arm. This raised the question whether breast cancer screening can cause substantial overdiagnoses. Chapter 11 shows that during a trial on breast cancer screening many more cancers can be diagnosed than in the control group while nevertheless there is only very limited extra incidence due to screening. That is because most of these cases are diagnosed earlier in women who would also be diagnosed with breast cancer at a later point in time. Screening for breast cancer therefore causes a substantial but mostly temporary increase of incidence.
Chapter 12 describes that the stage distribution of cancers that are detected at repeat breast cancer screenings is not much more favourable that that at first screenings while one would expect it to be much more favourable. Several mechanisms that may cause this difference between observed and differences in stage distribution are considered: little correlation between sojourn times in consecutive preclinical disease states, the presence of latent cancers, delay of symptomatic diagnosis due to participation with screening, and lower test sensitivity in repeat screenings. The first two mechanisms together can only explain part of the difference while it seems unlikely that reduction of sensitivity is occurring in so many screening projects. That leaves the possible mechanism of false reassurance. Also due to this finding an investigation into the occurrence of false reassurance in to the Dutch mass screening for breast cancer has been started.

The previous of this summary already presented conclusions with respect to the translation of evidence to policy making that has been clarified in this thesis. Chapter 13 discusses some issues in this translation that need further analysis: Remaining questions around uncertainty analysis are pointed out, and the balance of favourable and unfavourable health effects in screening for other cancer than that of the breast is considered. It is expected that the methodology of evaluating cancer screening will show considerable further development, particularly with respect to model based analysis, thus possibly leading to further rationalisation of decision making on cancer screening.

SAMENVATTING

Er liggen verschillende stappen tussen empirisch onderzoek en ondersteuning van besluitvorming over kankerscreening. De stappen van het verzamelen van de evidentie voor effectiviteit van screening, het evalueren van de onzekerheid rond schattingen van parameters en effecten, het opmaken van de balans tussen gunstige en ongunstige effecten en het schatten van de invloed van de lokale omstandigheden waaronder het bevolkingsonderzoek wordt uitgevoerd op de effecten van kankerscreening worden in dit proefschrift nader uitgewerkt aan de hand van een aantal voorbeelden. De meeste voorbeelden betreffen een toepassing van het micro-simulatiemodel Miscan dat speciaal ontwikkeld is voor de evaluatie van vroege opsporing van kanker.

verzamelen van evidentie

In 1993 werd een overzicht gepubliceerd van de vijf Zweedse gerandomiseerde trials op borstkankerscreening. Deze trials golden op dat moment als de belangrijkste bijdrage aan de evidentie voor borstkankerscreening. Hoofdstuk 2 beschrijft een gezamenlijke analyse van deze trials waarbij rekening wordt gehouden met de variatie in omstandigheden tussen de verschillende trials. De analyse is gericht op het afleiden van model parameters die nodig zijn voor het extrapoleren van de verzamelde evidentie naar andere screeningssituaties. Dit hoofdstuk gaat ook in op de vraag in hoeverre de waargenomen borstkankersterftereductie bij vrouwen die bij hun start van deelname aan de studie jonger dan 50 waren, verklaard kan worden door screening die later heeft plaatsgevonden wanneer zij ouder dan 50 jaar zijn. De conclusies waren dat de prognoseverbetering ten gevolge van ontdekking van borstkanker door screening groter is dan het resultaat van een eerdere schatting en dat de op dat moment gepubliceerde gegevens duidden op de mogelijkheid van een zeer geringe effectiviteit van screening bij vrouwen jonger dan 50 jaar. De laatste conclusie leidde tot publicatie van enkele gegevens die aangaven dat de veronderstellingen waaruit deze mogelijkheid werd afgeleid niet juist waren. Later is de evidentie voor effectiviteit van borstkankerscreening bij vrouwen jonger dan 50 jaar sterk toegenomen. De vraag naar de kosten-effectiviteit van borstkankerscreening bij jongere vrouwen moet echter nog voldoende worden beantwoord. De gepresenteerde methode van gezamenlijke analyse kan gezien worden als een aanvulling op de meta-analyse aanpak waarbij variatie in omstandigheden tussen verschillende trials en extrapolatie naar andere omstandigheden problematisch zijn, maar waarbij de statistische validiteit veel beter gewaarborgd is.

Behalve door gerandomiseerd onderzoek wordt de effectiviteit van screening ook wel geschat op basis van observationele gegevens. De meest aangewezen methode daarvoor is een case-control onderzoek met een aanpassing om rekening te houden met het feit dat na een diagnose van de ziekte geen screening meer plaatsvindt. Hoofdstuk 3 gaat na in hoeverre schattingen van kankersterftereductie ten gevolge van screening vertekend worden wanneer gebruik gemaakt wordt van een case-control onderzoeksopzet. Het blijkt dat op het eerste gezicht minimale verschillen in het matchen van cases en controls, en in het tellen van de blootstelling aan screening, kunnen leiden tot sterke vertekening. Zelfs bij de juiste matching en telling van blootstelling, resteert een belangrijke mate van vertekening doordat de aanpassing van de case-control onderzoeksopzet voor niet-deelname van prevalente gevallen, niet geheel accuraat is. Bovendien is er de mogelijkheid van vertekening ten gevolge van een associatie tussen deelname aan screening en risico op de ziekte. Case-control studies zouden derhalve maar in zeer beperkte mate een rol behoren te spelen bij de evaluatie van kankerscreening.

Hoofdstuk 4 beschrijft een onderzoek naar een alternatieve methode van het schatten van overleving na een diagnose van prostaatkanker waarbij in plaats van uit te gaan van een populatie gediagnosticeerde gevallen, uitgegaan wordt van sterfgevallen die gedurende een bepaalde periode hebben plaatsgevonden en waarbij vervolgens teruggezocht wordt naar een voorafgaande diagnose van prostaatkanker. Wanneer deze alternatieve methode wordt toegepast in een onveranderlijke situatie levert deze goede resultaten, maar wanneer de situatie in de loop van de studie verandert, bijvoorbeeld door vergrijzing van de bevolking of invoering van screening, ontstaat een belangrijke mate van vertekening.

In hoofdstuk 5 worden een aantal meer algemeen geaccepteerde methoden van schatting van overleving vergeleken. Het bleek dat de belangrijkste verschillen tussen schattingen voor netto overleving, althans bij dikke darmkanker en prostaatkanker, voortkomen uit de keuze voor ziekte-specifieke overleving of relatieve overleving (1,4% verschil in het overlevingspercentage bij dikke darmkanker en 6% bij prostaatkanker) en voor inclusie of exclusie van personen met meer dan één primaire kanker (ca. 4% verschil). De verschillen ten gevolge van andere varianten in methodiek bleken nog kleiner. Het verdient niettemin aanbeveling dat de keuze voor een bepaalde methode van schatten van overleving op mogelijke vertekening wordt beoordeeld.

evidentie en onzekerheid

In het Verenigd Koninkrijk waar vrouwen van 50-64 jaar oud iedere 3 jaar worden uitgenodigd voor borstkankerscreening, werd geconstateerd dat tijdens

het derde jaar na screening betrekkelijk veel intervalkankers voorkomen. Daarop rees de vraag of the screeningsinterval verkort moet worden van drie tot twee jaar. Hoofdstuk 6 gaat in op de vraag of het kosten-effectief is om het Britse borstkankscreeningsprogramma op deze wijze te intensiveren. De geschatte kosten per gewonnen levensjaar van een dergelijke intensivering zijn niet veel hoger dan die van het bestaande programma. Maar vanuit het oogpunt van kosten-effectiviteit is een verhoging van de laatste leeftijd van uitnodigingen van 64 naar 69 jaar evenzeer aan te bevelen.
Hoofdstuk 7 vergelijkt de kosten-effectiviteit van baarmoederhalskankerscreening door middel van cytologische uitstrijkjes met die door middel van directe visuele inspectie van de baarmoederhals. Deze vergelijking is met name interessant voor ontwikkelingslanden omdat directe inspectie veel minder kostbaar is en minder hoog ontwikkelde technische voorzieningen vergt. Echter, deze screeningsmethode mist meer afwijkingen en resulteert in een tamelijk groot percentage fout-positieven die aan nadere diagnostiek onderworpen moeten worden, zodat de verwachte totale kosten en beslag op technologie hoger zijn dan bij toepassing van het uitstrijkje bij een ongeveer even effectief screeningsprogramma. Echter, andere vormen van visuele inspectie met betere testeigenschappen, waarbij bijvoorbeeld de cervix eerst met wat verdund azijnzuur wordt behandeld of met behulp van een gynoscoop bekeken wordt, zijn mogelijk wel kosten-effectief.

afweging van gunstige en ongunstige effecten
Hoofdstuk 8 gaat na tot welke leeftijd van uitnodiging voor borstkankerscreening de gunstige gezondheidseffecten meer gewicht in de schaal leggen dan de ongunstige. Bij zeer pessimistische veronderstellingen blijkt deze leeftijdsgrens te liggen bij 75 jaar. Deze uitkomst gold als belangrijke grond voor de uitbreiding van de laatste uitnodiging voor het bevolkingsonderzoek van 69 tot 75 jaar. Bij meer waarschijnlijke veronderstellingen is deze uitbreiding vanuit het oogpunt van kosten-effectiviteit ook efficiënt.
In hoofdstuk 9 wordt overwogen of de ongunstige effecten van borstkankerscreening op hogere leeftijd beperkt kunnen worden door verlenging van het interval tussen uitnodigingen. Bij het verkorten van het interval worden vooral meer kankers gevonden met een betrekkelijk korte preklinische screen-detecteerbare duur. Bij deze kankers leidt vervroegde ontdekking vaak tot voorkomen van sterfte ten gevolge van deze kanker, en maar zelden tot extra incidentie van borstkanker. Daardoor nemen de ongunstige gezondheidseffecten in verhouding tot de gunstige minder sterk toe bij verkorting van het interval. De verhouding tussen gunstige en ongunstige gezondheidseffecten is dus geen aanleiding om op hogere leeftijd een langer screeningsinterval toe te passen. De kosten per gewonnen levensjaar zijn voor frequenter screening uiteraard wel hoger dan voor minder intensief screenen.

effectiviteit en omstandigheden
Hoofdstuk 10 beschrijft het belang van verschillende omstandigheden bij prospectieve evaluatie van de effecten en kosten van kankerscreening. Er wordt ingegaan op het grote belang van epidemiologische kenmerken zoals incidentieniveau, stadiumverdeling en overleving in de situatie zonder screening, de demografie, de kwaliteit van de screening, het screeningsschema, de screening die al heeft plaatsgevonden voor het moment van prospectieve evaluatie, de opkomst bij screening, de klinische praktijk buiten het screeningsprogramma, de kosten van verrichtingen, de balans tussen gunstige en ongunstige effecten en de mate van tijdvoorkeur. Hoewel de verschillen in omstandigheden van screening van groot belang kunnen zijn voor de keuze van een bepaald screeningsprogramma, lijken ze betrekkelijk weinig te maken te hebben met de verschillen in werkelijk uitgevoerde screening.

beantwoording van urgente vragen
De Malmö trial op borstkankerscreening liet zien dat tijdens de periode van screening ca. 25% meer borstkankers in de gescreende groep werden gevonden dan in de controle groep. Dit riep de vraag op of borstkankscreening een grote mate van overdiagnostiek veroorzaakt. Hoofdstuk 11 toont aan dat tijdens een trial op borstkankerscreening in de gescreende groep betrekkelijk veel meer kankers gevonden kunnen worden dan in de controle groep, terwijl toch sprake is van geringe extra incidentie ten gevolge van screening. Het gaat hierbij namelijk in de meeste gevallen om vervroeging van diagnose bij vrouwen bij wie zonder screening die diagnose ook, maar pas op een later tijdstip gesteld zou zijn. Screening op borstkanker veroorzaakt aldus een aanzienlijke tijdelijke verhoging van de incidentie.
Hoofdstuk 12 gaat in op het verschijnsel dat de stadiumverdeling van kankers die opgespoord worden bij vervolgscreeningen op borstkanker niet veel gunstiger blijkt te zijn dan bij eerste screeningen, terwijl de verwachting was dat deze veel gunstiger zou zijn. Er worden verschillende mechanismen voor het verschil tussen waarneming en verwachting overwogen: weinig correlatie tussen verblijftijden in verschillende opeenvolgende preklinische ziektestadia, de aanwezigheid van latente kankers, uitstel van diagnostiek op basis van symptomen ten gevolge van deelname aan screening, en in vervolgonderzoeken afgenomen sensitiviteit

van de screening. De eerste twee van deze mechanismen kunnen gezamenlijk slechts een deel van het verschil verklaren en het werd onwaarschijnlijk geacht dat afname van sensitiviteit in zo veel screeningsprojecten een rol speelde. Zodoende resteert het mogelijke mechanisme van 'onterechte geruststelling'. Mede naar aanleiding van deze bevinding is nu een onderzoek gestart naar onterechte geruststelling door het bevolkingsonderzoek naar borstkanker.

In het voorgaande zijn al conclusies gepresenteerd over vraagstellingen rond vertaling van evidentie naar beleid waarover in het proefschrift meer helderheid is verkregen. In hoofdstuk 13 worden een aantal vraagstukken rond deze vertaalslag besproken die verdere analyse behoeven: Er worden enige resterende vragen rond onzekerheidsanalyse aangereikt, en de balans van de gunstige en ongunstige gezondheidseffecten bij andere kankerscreening dan die op borstkanker wordt beproken. Naar verwachting zal de methodologie voor de evaluatie van kankerscreening zich nog aanmerkelijk verder ontwikkelen, met name op het gebied van de modelmatige analyse van problemen, waarmee een verdere rationalisering van de besluitvorming rond kankerscreening kan worden bereikt.

REFERENCES

Abrahamsson, A., H. O. Adami, A. Taube, K. Kim, M. Zelen, and M. Kulldorff. 1996. Re: Long-term survival and mortality in prostate cancer treated with noncurative intent [letter; comment]. J Urol 155 (1):296-298.

Anderson, T. J., J. Lamb, P. Donnan, F. E. Alexander, A. Huggins, and et al. 1991. Comparative pathology of breast cancer in a randomised trial of screening. Br J Cancer 64 (1):108-113.

Andersson, I. 1984. Breast cancer screening in Malmö. Recent Results Cancer Res 90:114-116.

Andersson, I., K. Aspegren, L. Janzon, T. Landberg, K. Lindholm, and et al. 1988. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ 297:943-948.

Aristizabal, N., C. Cuello, P. Correa, T. Collazos, and W. Haenszal. 1984. The impact of vaginal cytology on cervical cancer risks in Cali, Colombia. International J. of Cancer 34:5-9.

Aus, G., J. Hugosson, and L. Norlen. 1995. Long-term survival and mortality in prostate cancer treated with noncurative intent [see comments]. J Urol 154 (2 Pt 1):460-465.

Baines, C. J. 1994. The Canadian National Breast Screening Study: a perspective on criticisms. Ann Intern Med 120:326-334.

Baines, C. J., and A. B. Miller. 1997. Mammography versus clinical examination of the breasts. Monogr Natl Cancer Inst (22):125-129.

Baldauf, J. J., M. Dreyfus, J. Ritter, P. Meyer, and E. Philippe. 1997. Cervicography. Does it improve cervical cancer screening? Acta Cytol 41 (2):295-301.

BCDDP. 1979. Report of the Working Group to Review the National Cancer Institute-American Cancer Society Breast Cancer Detection Demonstration Projects. J Natl Cancer Inst 62 (3):639-709.

Bedell, M. B., M. E. Wood, D. C. Lezotte, S. M. Sedlacek, and M. M. Orleans. 1995. Delay in diagnosis and treatment of breast cancer: implications for education. J Cancer Educ 10 (4):223-228.

Beemsterboer, P. M. M. 1999. Evaluation of screening programmes. Studies on breast cancer and prostate cancer [dissertation], Department of Public Health, Erasmus University Rotterdam, Rotterdam.

Beemsterboer, P. M. M., H. J. de Koning, P. G. Warmerdam, R. Boer, E. Swart, Dierks M.L., and et al. 1994. Prediction of the effects and costs of breast-cancer screening in Germany. Int J Cancer 58:623-628.

Berget, A. 1979. Influence of population screening on morbidity and mortality of cancer of the uterine Cervix in Maribo Amt. Dan Med Bull 26 (2):91-100.

Berrino, F. 1993. Personal communication on screening exposure and the DPCP.

Berrino, F., G. Gatta, M. D'Alto, P. Crosignani, and E. Riboli. 1984. Use of case-control studies in evaluation of screening programmes. UICC Technol Rep Ser 78:24-43.

Bhargava, V. L., K. Verma, R. Sharma, S. Batra, and P. N. Anandalakshmy. 1993. A hospital-based study on the use of paramedical personnel for clinical downstaging of cancer cervix. Indian J Med Res 98:65-68.

Black, R. J., F. Bray, J. Ferlay, and D. M. Parkin. 1997. Cancer incidence and mortality in the European Union: cancer registry data and estimates of national incidence for 1990 [see comments] [published erratum appears in Eur J Cancer 1997 Dec;33(14):2440]. Eur J Cancer 33 (7):1075-1107.

Blanks, R. G., S. M. Moss, and M. G. Wallis. 1996. A comparison of two view and one view mammography in the detection of small invasive cancers: results from the National Health Service breast screening programme. J Med Screen 3 (4):200-203.

Blettner, M., W. Sauerbrei, B. Schlehofer, T. Scheuchenpflug, and C. Friedenreich. 1999. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol 28 (1):1-9.

Boer, R., H. J. de Koning, P. M. M. Beemsterboer, P. G. Warmerdam, and F. H. Schröder. 1997. A comparison of disease specific survival of patients who died of and who had newly diagnosed prostate cancer. J Urol 157 (5):1768-1771; discussion 1771-1772.

Boer, R., H.J. de Koning, A. Threfall, P. Warmerdam, A. Street, and et al. 1998. Cost-effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study [see comments]. BMJ 317 (7155):376-379.

Boer, R., H. J. de Koning, G. J. van van Oortmarssen, and P.J. van der Maas. 1995. In search of the best upper age limit for breast cancer screening. Eur J Cancer 31A (12):2040-2043.

Boer, R., G. J. van Oortmarssen, and H.J. de Koning. 1995. Cost-effectiveness of mammography screening.(letter + reply Rosenquist and Lindfors). JAMA 275 (2):111-112.

Boer, R., P. Warmerdam, and H.J. de Koning. 1994. Extra incidence caused by mammographic screening [letter]. Lancet 343:979.

Boyes, D. A., B. Morrison, E. G. Knox, G. J. Draper, and A. B. Miller. 1982. A cohort study of cervical cancer screening in British Columbia. Clin Invest Med 5:1-29.

Brekelmans, C. T., J. M. van Gorp, P. H. Peeters, and H. J. Collette. 1996. Histopathology and growth rate of interval breast carcinoma. Characterization of different subgroups. Cancer 78 (6):1220-1228.

Brown, B. W., C. Brauner, and M. C. Minnotte. 1993. Noncancer deaths in white adult cancer patients. J Natl Cancer Inst 85 (12):979-987.
Brown, M. L., and L. Fintor. 1993. Cost-effectiveness of breast cancer screening: preliminary results of a systematic review of the literature. Breast Cancer Research Treat 25 (2):113-118.
Burns, P. E. 1978. False-negative mammograms delay diagnosis of breast cancer. N Engl J Med 299 (4):201-202.
Cairns, J. A., and M. M. van der Pol. 1997. Saving future lives. A comparison of three discounting models. Health Econ 6 (4):341-350.
Cantor, S. B. 1994. Cost-effectiveness analysis, extended dominance, and ethics: a quantitative assessment. Med Decis Making 14 (3):259-265.
Caplan, L. S., and K. J. Helzlsouer. 1992. Delay in breast cancer: a review of the literature. Public Health Rev 20 (3-4):187-214.
Caplan, L. S., K. J. Helzlsouer, S. Shapiro, M. N. Wesley, and B. K. Edwards. 1996. Reasons for delay in breast cancer diagnosis. Prev Med 25 (2):218-224.
CBS. 1992. Life tables, 1991 and 1987-1991. Mndstat Bevolk (CBS) 11 (12):41-44.
CBS. 1996. Male deaths by cause of death and age and average male population, 1994. Mndstat Bevolk (CBS) 15 (2):28-29.
Celentano, D. D., A. C. Klassen, C. S. Weisman, and N. B. Rosenshein. 1988. Cervical cancer screening practices among older women: results from the Maryland Cervical Cancer Case-Control Study. J Clin Epidemiol 41 (6):531-541.
Chamberlain, J. 1993. Firmer evidence on the value of breast screening - the Swedish overview. Eur J Cancer 29A (13):1804-1805.
Chamberlain, J., S. M. Moss, A. E. Kirkpatrick, M. Michell, and L. Johns. 1993. National Health Service breast screening programme results for 1991-2. [publisched erratum appears in BMJ 1993;307:543, see comments BMJ 1993;307:353-356]. BMJ 307:353-356.
Chang, H. R., B. Cole, and K. I. Bland. 1997. Nonpalpable breast cancer in women aged 40-49 years: a surgeon's view of benefits from screening mammography. Monogr Natl Cancer Inst (22):145-149.
Chen, H. H., S. W. Duffy, L. Tabár, and N. E. Day. 1997. Markov chain models for progression of breast cancer. Part II: prediction of outcomes for different screening regimes. J Epidemiol Biostat 2 (1):25-35.
Chessa, A.G., M.E. van den Akker-van Marle, G.J. van Oortmarssen, R. Boer, M. van Ballegooijen, and J.D.F. Habbema. submitted. Uncertainty analysis of the const-effectiveness of cervical cancer screening policies using the Miscan microsimulation program. .
Chodak, G. W., R. A. Thisted, G. S. Gerber, J. E. Johansson, J. Adolfsson, and et al. 1994. Results of conservative management of clinically localized prostate cancer. N Engl J Med 330 (4):242-248.
Clarke, E. A., and T. W. Anderson. 1979. Does screening by "Pap" smears help prevent cervical cancer? Lancet 2 (8132):1-4.
Collette, H. J. A., J. J. Rombach, N. E. Day, and F. de Waard. 1984. Evaluation of screening for breast cancer in a non-randomised study (The DOM Project) by means of a case-control study. Lancet 1 (8388):1224-1226.
Costanza, M. E. 1992. Breast cancer screening in older women. Cancer 69 (7 Supp):1925-1931.
Cox, B. 1997. Variation in the effectiveness of breast screening by year of follow-up. Monogr Natl Cancer Inst (22):69-72.
Cox, D.C., and P. Baybutt. 1981. Methods for uncertainty analysis: a comparitive survey. Risk Analysis 1:251-258.
Cronin, K. A., D. L. Weed, R. J. Connor, and P. C. Prorok. 1998. Case-control studies of cancer screening: theory and practice. J Natl Cancer Inst 90 (7):498-504.
Davey Smith, G., and M. Egger. 1998. Meta-analysis. Unresolved issues and future developments. Bmj 316 (7126):221-225.
Davey Smith, G., M. Egger, and A. N. Phillips. 1997. Meta-analysis. Beyond the grand mean? BMJ 315 (7122):1610-1614.
Day, N.E., S.D. Walter, L. Tabár, C.J.G. Fagerberg, and H.J.A. Collette. 1988. The sensitivity and lead time of breast cancer screening: a comparison of the results of different studies. In Screening for breast cancer, edited by N. E. Day and A. B. Miller. Toronto: Hans Huber Publishers.
de Haes, J, C.J.M., H. J. de Koning, G. J. van Oortmarssen, H. M. E. van Agt, A. E. de Bruyn, and P. J. van der Maas. 1991. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer 49:538-544.
de Koning, H.J. 1993. The effects and costs of breast cancer screening [dissertation], Erasmus University Rotterdam, Rotterdam.
de Koning, H. J., R. Boer, P. G. Warmerdam, P. M. M. Beemsterboer, and P. J. van der Maas. 1995a. Quantitative interpretation of age-specific mortality reductions from the Swedish Breast Cancer-Screening Trials. J Natl Cancer Inst 87 (16):1217-1223.
de Koning, H. J., J. Fracheboud, R. Boer, A. L. Verbeek, H. J. Collette, J. H. Hendriks, B. M. van Ineveld, A. E. de Bruyn, and P. J. van der Maas. 1995b. Nation-wide breast cancer screening in The Netherlands: support for breast-cancer mortality reduction. National Evaluation Team for Breast Cancer Screening (NETB). Int J Cancer 60 (6):777-780.
de Koning, H.J., J. A. van Dongen, and P.J. van der Maas. 1994. Changes in use of breast-conserving therapy in years 1987-2000. Br J Cancer 70:1165-1170.

de Koning, H. J., B. M. van Ineveld, G. J. van Oortmarssen, J.C.J.M. de Haes, H. J. A. Collette, and et al. 1991. Breast cancer screening and cost-effectiveness; policy alternatives, quality of life considerations and the possible impact of uncertain factors. Int J Cancer 49:531-537.

De Sutter, Ph, M. Coibion, M. Vosse, D. Hertens, F. Huet, F. Wesling, M. Wayembergh, C. Bourdon, and Ph Autier. 1998. A multicentre study comparing cervicography and cytology in the detection of cervical intraepithelial neoplasia. Br J Obstet Gynaecol 105 (6):613-620.

Drummond, M.F., B. O'Brien, G.L. Stoddart, and G.W. Torrance. 1997. Methods for the economic evaluation of health care programmes [second edition]. Oxford: Oxford University Press.

Duffy, S. W., L. Tabár, G. Fagerberg, A. Gad, O. Gröntoft, and et al. 1991. Breast screening, prognostic factors and survival - results from the Swedish two county study. Br J Cancer 64:1133-1138.

Eastman, P. 1997. NCI adopts new mammography screening guidelines for women [news]. J Natl Cancer Inst 89 (8):538-540.

Ebeling, K., and P. Nischan. 1987. Screening for lung cancer--results from a case-control study. Int J Cancer 40 (2):141-144.

Eckhardt, S., F. Badellino, and G. P. Murphy. 1994. UICC meeting on breast-cancer screening in pre-menopausal women in developed countries. Geneva, 29 September - 1 October 1993. Int J Cancer 56:1-5.

Eddy, D. M. 1990. Screening for cervical cancer [see comments]. Ann Intern Med 113 (3):214-226.

Ederer, F., L.M. Axtell, and S.J. Cutler. 1961. The relative survival rate: a statistical methodology. Natl Cancer Inst Monogr 6:101-121.

Efron, B., and R. Tibshirani. 1993. An introduction to the bootstrap. New York: Chapman & Hall.

Ekelund, RB, and RD Tollison. 1986. Economics. Boston: Little, Brown and Company.

Elwood, J. M., B. Cox, and A. K. Richardson. 1993. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993 Mar 5;Doc No 34:[295 words; 2 paragraphs] and 1994 Mar 31;Doc No 121:[385 words; 4 paragraphs]] [see comments]. Online J Curr Clin Trials Doc No 32:[23,227 words; 195 paragraphs].

Elwood, J. M., and W. P. Moorehead. 1980. Delay in diagnosis and long-term survival in breast cancer. Br Med J 280:1291-1294.

Estève, J., E. Benhamou, and L. Raymond. 1994. Statistical methods in cancer research. Vol. VI. Lyon: WHO/IARC.

Fagerberg, G., L. Baldetorp, O. Grontoft, B. Lundstrom, J. C. Manson, and B. Nordenskjold. 1985. Effects of repeated mammographic screening on breast cancer stage distribution: Results from a randomised study of 92 934 women in a Swedish county. Acta Radiol 24 (6):465-473.

Feig, S. A., and R. E. Hendrick. 1997. Radiation risk from screening mammography of women aged 40-49 years. Monogr Natl Cancer Inst (22):119-124.

Feuer, E. J., R. M. Merrill, and B. J. Hankey. 1999. Cancer surveillance series: Interpreting trends in prostate cancer - Part II: Cause of death misclassification and the recent rise and fall in prostate cancer mortality. J Natl Cancer Inst 91 (12):1025-1032.

Fletcher, S. W. 1997. Breast cancer screening among women in their forties: an overview of the issues. Monogr Natl Cancer Inst (22):5-9.

Fletcher, S. W., W. Black, R. Harris, B. K. Rimer, and S. Shapiro. 1993a. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 85 (20):1644-1656.

Fletcher, S. W., W. Black, R. Harris, B. K. Rimer, and S. Shapiro. 1993b. Report of the International Workshop on Screening for Breast Cancer [see comments]. J Natl Cancer Inst 85 (20):1644-1656.

Forrest, A. P., and F. E. Alexander. 1995. A question that will not go away: at what age shoud mammographic screening begin? J Natl Cancer Inst 87 (16):1195-1199.

Forrest, A.P.M. 1987. Report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a Working Group chaired by Sir Patrick Forrest. London: Her Majesties Stationary Office.

Forrest, A. P. M., and R. J. Aitken. 1990. Mammography screening for breast cancer. Annu Rev Med 41:117-132.

Forrest, P. 1986. Breast cancer screening. Report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a working group chaired by Prof. Sir Patrick Forrest. London: Her Majesty's Stationery Office.

Fracheboud, J., R. Boer, B. M. van Ineveld, P. M. M. Beemsterboer, A. L. M. Verbeek, and et al. 1997. Landelijke evaluatie van bevolkingsonderzoek naar borstkanker in Nederland - VI, Het zesde evaluatie-rapport. Edited by R. Erasmus Universiteit, U. Universiteit and N. Kath. universiteit. Rotterdam: Letb.

Frankel, S. D., E. A. Sickles, B. N. Curpen, R. A. Sollitto, S. H. Ominsky, and H. B. Galvin. 1995. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. AJR Am J Roentgenol 164 (5):1107-1109.

Friedman, G. D., R. A. Hiatt, Quesenberry Cp, Jr., J. V. Selby, and N. S. Weiss. 1995. Problems in assessing screening experience in observational studies of screening efficacy: example of urinalysis screening for bladder cancer. J Med Screen 2 (4):219-223.

Frisch, L. E., F. H. Milner, and D. G. Ferris. 1994. Naked-eye inspection of the cervix after acetic acid application may improve the predictive value of negative cytologic screening [see comments]. J Fam Pract 39 (5):457-460.

Frischbier, H.J., B.P. Robra, W. Hoeffken, and M.L. Dierks. 1994. Mammographie in der Krebsfrueherkennung. Stuttgart: Ferdinand Enke Verlag.

Frisell, J., G. Eklund, L. Hellström, U. Glas, and A. Somell. 1989. The Stockholm breast cancer screening trial - 5 year results and stage at discovery. Breast Cancer Res Treat 13:79-87.

Frisell, J., U. Glas, L. Hellström, and A. Somell. 1986. Randomized mammographic screening for breast cancer in Stockholm. Breast Cancer Res Treat 8:45-54.

Gaffikin, L., P.D. Blumenthal, J. McGrath, and Z.M. Chirenje. 1999. Visual inspection with acetic acid for cervical-cancer screening: test qualities in a primary-care setting. University of Zimbabwe/JHPIEGO Cervical Cancer Project. Lancet 353 (9156):869-873.

George, N. J. R. 1988. Natural history of localised prostatic cancer managed by conservative therapy alone. Lancet i:494-497.

Gill, T. M., and R. I. Horwitz. 1995. Evaluating the efficacy of cancer screening: clinical distinctions and case-control studies [see comments]. J Clin Epidemiol 48 (2):281-292.

Glasziou, P., and L. Irwig. 1997. The quality and interpretation of mammographic screening trials for women ages 40-49. Monogr Natl Cancer Inst (22):73-77.

Gøtzsche, P. C., and O. Olsen. 2000. Is screening for breast cancer with mammography justifiable? Lancet 355 (9198):129-134.

Grönberg, H. 1995. Prostate cancer, epidemiological studies [dissertation], Umea University, Umea.

Gullberg, B., I. Andersson, L. Janzon, and J. Ranstam. 1991. Screening mammography. Lancet 337:244.

Gustafsson, L., J. Ponten, R. Bergstrom, and H. O. Adami. 1997. International incidence rates of invasive cervical cancer before cytological screening. Int J Cancer 71 (2):159-165.

Habbema, J. D.F., S. J. de Vlas, A. P. Plaisier, and G. J. van Oortmarssen. 1996. The microsimulation approach to epidemiologic modeling of helminthic infections, with special reference to schistosomiasis. Am J Trop Med Hyg 55 (5 Suppl):165-169.

Habbema, J. D.F., J.Th.N. Lubbe, G. J. van Oortmarssen, and P. J. van der Maas. 1987. A simulation approach to cost-effectiveness and cost-benefit calculations of screening for the early detection of disease. Eur J Oper Res 29:159-166.

Habbema, J. D.F., G. J. van Oortmarssen, J.Th.N. Lubbe, and P. J. van der Maas. 1985. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed 20 (1):79-93.

Habbema, J. D.F., G. J. van Oortmarssen, D. J. van Putten, J. T. Lubbe, and P.J. van der Maas. 1986. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York Study. J Natl Cancer Inst 77 (2):317-320.

Hakama, M., K. Holli, J. Isola, O. P. Kallioniemi, A. Kärkkäinen, and et al. 1995. Aggressiveness of screen-detected breast cancers. Lancet 345:221-224.

Harris, R. 1997. Variation of benefits and harms of breast cancer screening with age. Monogr Natl Cancer Inst (22):139-143.

Hosek, R. S., W. D. Flanders, and A. J. Sasco. 1996. Bias in case-control studies of screening effectiveness. Am J Epidemiol 143 (2):193-201.

Isaacs, D., and D. Fitzgerald. 1999. Seven alternatives to evidence based medicine. BMJ 319 (7225):1618.

Johansson, J. E., H. O. Adami, S. O. Andersson, R. Bergström, L. Holmberg, and U. B. Krusemo. 1992. High 10-year survival rate in patients with early, untreated prostatic cancer. JAMA 267 (16):2191-2196.

Kaplan, E.L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457-481.

Koopmanschap, M.A. 1994. Complementary analyses in economic evaluation of health care [dissertation], Erasmus University Rotterdam, Rotterdam.

Koopmanschap, M. A., G. J. van Oortmarssen, H. M.E. van Agt, M. van Ballegooijen, J. D.F. Habbema, and J.Th.N. Lubbe. 1990. Cervical-cancer screening: attendance and cost-effectiveness. Int J Cancer 45 (3):410-415.

Kopans, D. B. 1997a. Canadian National Breast Screening Study [letter]. Lancet 350 (9080):810-810.

Kopans, D. B. 1997b. NBSS: opportunity to compromise the process. CMAJ 157 (3):247-248.

Kopans, D. B. 1997c. An overview of the breast cancer screening controversy. Monogr Natl Cancer Inst (22):1-3.

Libstug, A. R., V. Moravan, and S. E. Aitken. 1998. Results from the Ontario breast screening program, 1990-1995. J Med Screen 5 (2):73-80.

Loeve, F., R. Boer, G. J. van Oortmarssen, M. van Ballegooijen, and J. D.F. Habbema. 1999. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 32:13-33.

MacGregor, J. E., S. M. Moss, D. M. Parkin, and N. E. Day. 1985. A case-control study of cervical cancer screening in north east Scotland. BMJ 290:1543-1546.

McCann, J., S. Wait, B. Seradour, and N. Day. 1997. A comparison of the performance and impact of breast cancer screening programmes in East Anglia, U.K. and Bouches du Rhône, France. Eur J Cancer 33 (3):429-435.

Megevand, E., L. Denny, K. Dehaeck, R. Soeters, and B. Bloch. 1996. Acetic acid visualization of the cervix: an alternative to cytologic screening. Obstet Gynecol 88 (3):383-386.

Michaelson, J. S., E. Halpern, and D. B. Kopans. 1999. Breast cancer: computer simulation method for estimating optimal intervals for screening. Radiology 212 (2):551-560.

Miller, A. B., C. J. Baines, T. To, and C. Wall. 1992. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. Can Med Assoc J 147 (10):1459-1476.

Miller, A. B., T. To, C. J. Baines, and C. Wall. 1997. The Canadian National Breast Screening Study: update on breast cancer mortality. Monogr Natl Cancer Inst (22):37-41.

Morgan, M.G., and M. Henrion. 1990. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. New York: Cambridge University Press.

Morrison, A. S. 1982. Case definition in case-control studies of the efficacy of screening. Am J Epidemiol 115 (1):6-8.

Moss, S. M. 1991. Case-control studies of screening. Int J Epidemiol 20 (1):1-6.

Nandakumar, A., N. Anantha, and T. C. Venugopal. 1995. Incidence, mortality and survival in cancer of the cervix in Bangalore, India. Br J Cancer 71 (6):1348-1352.

National Cancer Institute, DCPC, Surveillance Program, Cancer Statistics Branch,. 1997. Surveillance, Epidemiology, and End Results (SEER) Program public use CD-ROM (1973-1994), released May 1997, based on the August 1996 Submission. Bethesda, MD: National Cancer Institute.

Neave, L. M., B. H. Mason, and R. G. Kay. 1990. Does delay in diagnosis of breast cancer affect survival? Breast Cancer Res Treat 15 (2):103-108.

NeCaRe. 1991. Incidence of cancer in the Netherlands 1991. Utrecht: NeCaRe.

Nene, B. M., S. Deshpande, K. Jayant, A. M. Budukh, P. S. Dale, D. A. Deshpande, A. S. Chiwate, S. G. Malvi, S. Deokar, D. M. Parkin, and R. Sankaranarayanan. 1996. Early detection of cervical cancer by visual inspection: a population-based study in rural India. Int J Cancer 68 (6):770-773.

Nyström, L, and L-G. Larsson. 1993. Re: Breast cancer screening with mammography [letter]. Lancet 341:1531-1532.

Nyström, L., L. E. Rutqvist, S. Wall, A. Lindgren, M. Lindqvist, and et al. 1993. Breast cancer screening with mammography: overview of Swedish randomised trials. Published erratum appears in Lancet 1993 Nov 27;342(8883):1372. Lancet 341 (8851):973-978.

Oshima, A., N. Hirata, T. Ubukata, K. Umeda, and I. Fujimoto. 1986. Evaluation of a mass screening program for stomach cancer with a case- control study design. Int J Cancer 38 (6):829-833.

Otten, J. D., J. A. van Dijck, P. G. Peer, H. Straatman, A. L. Verbeek, and M. Mravunac. 1996. Long term breast cancer screening in Nijmegen, The Netherlands: the nine rounds from 1975-92. J Epidemiol Community Health 50 (3):353-358.

Paci, E., R. Boer, M. Zappa, H. J. de Koning, G. J. van Oortmarssen, E. Crocetti, and et al. 1995. A model-based prediction of the impact on reduction in mortality by a breast cancer screening programme in the city of Florence, Italy. Eur J Cancer 31A (3):348-353.

Palli, D., M. Rosselli del Turco, E. Buiatti, S. Carli, S. Ciatto, and L. Toscani. 1986. A case-control study of the efficacy of a non-randomized breast cancer screening program in Florence (Italy). Int J Cancer 38:501-504.

Parkin, D. M., C. S. Muir, S. L. Whelan, Y. T. Gao, J. Ferlay, and J. Powell. 1992. Cancer incidence in five continents. Vol. VI. Lyon: WHO/IARC.

Peer, P. G. M., J.A.A.M. van Dijck, J.H.C.L. Hendriks, R. Holland, and A. L. M. Verbeek. 1993. Age-dependent growth rate of primary breast cancer. Cancer 71 (11):3547-3551.

Peer, P. G. M., R. Holland, J.H.C.L. Hendriks, M. Mravunac, and A. L. M. Verbeek. 1994. Age-specific effectiveness of the Nijmegen population-based breast cancer-screening program: assessment of early indicators of screening effectiveness [see comments]. J Natl Cancer Inst 86 (6):436-441.

Percy, C., V.V. Holten, and C. Muir. 1990. International classification of diseases for oncology [2nd edition]. Geneva: WHO.

Percy, C. L., B. A. Miller, and L. A. Gloeckler Ries. 1990. Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in cancer mortality. Ann N Y Acad Sci 609:87-97; discussion 97-99.

Porta, M., M. Gallén, N. Malats, and J. Planas. 1991. Influence of "diagnostic delay" upon cancer survival: an analysis of five tumour sites. J Epidemiol Community Health 45 (3):225-230.

Registrar General of India. 1995. SRS Based Abridged Life Tables 1988-92. Occasional Paper No.4 of 1995. New Delhi: Office of the Registrar General.

Richards, M. A., S. Braysher, W. M. Gregory, and R. D. Rubens. 1993. Advanced breast cancer: use of resources and cost implications. Br J Cancer 67 (4):856-860.

Roberts, M. M., F. E. Alexander, T. J. Anderson, U. Chetty, P. T. Donnan, and et al. 1990. Edinburgh trial of screensing for breast cancer: mortality at seven years. Lancet 335:241-246.

Russell, L. B., M. R. Gold, J. E. Siegel, N. Daniels, and M. C. Weinstein. 1996. The role of cost-effectiveness analysis in health and medicine. Panel on Cost-Effectiveness in Health and Medicine [see comments]. JAMA 276 (14):1172-1177.

Salzmann, P., K. Kerlikowske, and K. Phillips. 1997. Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. Ann Intern Med 127 (11):955-965.

Sankaranarayanan, R., B. Shyamalakumary, R. Wesley, N. Sreedevi Amma, D. M. Parkin, and M. K. Nair. 1999. Visual inspection with acetic acid in the early detection of cervical cancer and precursors [letter]. Int J Cancer 80 (1):161-163.

Sankaranarayanan, R., R. Wesley, T. Somanathan, N. Dhakad, B. Shyamalakumary, and N. S. Amma. 1998. Visual inspection of the uterine cervix after the application of acetic acid in the detection of cervical carcinoma and its precursors. Cancer 83 (10):2150-2156.

SAS. 1991. SAS/STAT Software: the PHREG Procedure, Version 6, Cary NC: SAS Institute Inc.

Sasco, A. J., N. E. Day, and S. D. Walter. 1986. Case-control studies for the evaluation of screening. J Chron Dis 39 (5):399-405.

Schrijvers, C. T., J. W. W. Coebergh, L. H. van der Heijden, and J. P. Mackenbach. 1995. Socioeconomic variation in cancer survival in the southeastern Netherlands, 1980-1989. Cancer 75 (12):2946-2953.

Scottish Cancer Therapy Network. 1996. Scottish breast cancer audit 1987 and 1993. Report to the Chief Scientist and CRAG: Scottish Cancer Therapy Network.

Sehgal, A., V. Singh, S. Bhambhani, and U. K. Luthra. 1991. Screening for cervical cancer by direct inspection. Lancet 338 (8762):282.

Selby, J. V., G. D. Friedman, C. P. Quesenberry, Jr., and N. S. Weiss. 1992. A case-control study of screening sigmoidoscopy and mortality from colorectal cancer. N Engl J Med 326 (10):653-657.

Shapiro, S., P. Strax, L. Venet, and W. Venet. 1974. Changes in 5-year breast cancer mortality in a breast cancer screening program. In Seventh National Cancer Conference Proceedings: American Cancer Society.

Shapiro, S., W. Venet, P. Strax, L. Venet, and R. Roeser. 1982. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 69 (2):349-355.

Shapiro, S., W. Venet, P. Strax, L. Venet, and R. Roeser. 1985. Selection, follow-up, and analysis in the Health Insurance Plan Study: a randomized trial with breast cancer screening. Natl Cancer Inst Monogr 67:65-74.

Siegel, J. E., M. C. Weinstein, L. B. Russell, and M. R. Gold. 1996. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine [see comments]. JAMA 276 (16):1339-1341.

Singh, V., A. Sehgal, and U. K. Luthra. 1992. Screening for cervical cancer by direct inspection. BMJ 304:534-535.

Smith, R. A. 1995. Screening women aged 40-49 years: where are we today? J Natl Cancer Inst 87 (16):1198-1200.

Street, A.D., J. Posnett, A.G. Threlfall, C.B.J. Woodman, E. Twelves, E.H.I. Friedman, and et al. 1996. Economic evaluation of proposed changes to the breast screening programme. York: Health Economics Consortium: University of York.

Sujathan, K., S. Kannan, K. R. Pillai, A. Mathew, M. Joseph, B. Symalakumari, and M. K. Nair. 1995. Implications of gynaecological abnormalities in pre-selection criteria for cervical screening: preliminary evaluation of 3602 subjects in south India. Cytopathology 6 (2):75-87.

Tabár, L. 1996. Re: quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials.(letter + response Koning de, et al). J Natl Cancer Inst 88 (1):52-53.

Tabár, L., E. Akerlund, and A. Gad. 1984. Five-year experience with single-view mammography randomized controlled screening in Sweden. Recent Results Cancer Res 90:105-513.

Tabár, L., C. J. Fagerberg, A. Gad, L. Baldetorp, L. H. Holmberg, O. Grontoft, U. Ljungquist, B. Lundstrom, J. C. Manson, and G. Eklund. 1985. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1:829-832.

Tabár, L., C.J.G. Fagerberg, and N.E. Day. 1988. Evaluation of the results. In Screening for breast cancer, edited by N. E. Day and A. B. Miller. Toronto: Hans Huber Publishers.

Tabár, L., G. Fagerberg, H. H. Chen, S. W. Duffy, C. R. Smart, and et al. 1995. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 75 (10):2507-2517.

Tabár, L., G. Fagerberg, N. E. Day, and L. Holmberg. 1987. What is the optimum interval between mammographic screening examinations?- An analysis based on the latest results of the Swedish two-county breast cancer screening trial. Br J Cancer 55:547-551.

Tabár, L., G. Fagerberg, S. W. Duffy, and N. Day. 1989. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Community Health 43 (2):107-114.

Tabár, L., G. Fagerberg, S. W. Duffy, N. E. Day, A. Gad, and O. Gröntoft. 1992. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 30 (1):187-210.

Thurfjell, E. L., and J. A.Å. Lindgren. 1994. Population-based mammography screening in Swedish clinical practice: prevalence and incidence sreening in Uppsala County. Radiology 193 (2):351-357.

van Ballegooijen, M. 1998a. Comparing the predictions of a cervical cancer natural history model with incidence and mortality trends after the introduction of screening. In Effects and costs of cervical cancer screening [dissertation], edited by M. van Ballegooijen. Rotterdam: Erasmus Universiteit Rotterdam.

van Ballegooijen, M. 1998b. Effects and costs of cervical cancer screening [dissertation], Erasmus University Rotterdam, Rotterdam.

van Ballegooijen, M., J. D.F. Habbema, G. J. van Oortmarssen, M. A. Koopmanschap, J.Th.N. Lubbe, and H. M. E. van Agt. 1992a. Preventive pap-smears: balancing costs, risks and benefits. Br J Cancer 65:930-933.

van Ballegooijen, M., M. A. Koopmanschap, and J. D.F. Habbema. 1995. The management of cervical intra-epithelial neoplasia (CIN): extensiveness and costs in The Netherlands. Eur J Cancer 31A (10):1672-1676.

van Ballegooijen, M., M. A. Koopmanschap, A. J. Subandono, A. J. S. Tjokrowardojo, and G. J. van Oortmarssen. 1992b. Care and costs for advanced cervical cancer. Eur J Cancer 28A (10):1703-1708.

van Ballegooijen, M., M. A. Koopmanschap, G. J. van Oortmarssen, J. D.F. Habbema, J.Th.N. Lubbe, and H. M. van Agt. 1990. Diagnostic and treatment procedures induced by cervical cancer screening. Eur J Cancer 26 (9):941-945.

van Ballegooijen, M., M. E. van den Akker-van Marle, P. G. Warmerdam, C. J. Meijer, J. M. Walboomers, and J. D.F Habbema. 1997. Present evidence on the value of HPV testing for cervical cancer screening: a model-based exploration of the (cost-)effectiveness. Br J Cancer 76 (5):651-657.

van den Akker-van Marle, M. E., C. M. Reep-van den Bergh, R. Boer, A. Del Moral, N. Ascunce, and H. J. de Koning. 1997. Breast cancer screening in Navarra: interpretation of a high detection rate at the first screening round and a low rate at the second round. Int J Cancer 73 (4):464-469.

van Ineveld, B. M., G. J. van Oortmarssen, H. J. de Koning, R. Boer, and P. J. van der Maas. 1993. How cost-effective is breast cancer screening in different EC countries? Eur J Cancer 29A (12):1663-1668.

Van Le, L., F. F. Broekhuizen, R. Janzer-Steele, M. Behar, and T. Samter. 1993. Acetic acid visualization of the cervix to detect cervical dysplasia. Obstet Gynecol 81 (2):293-295.

van Oortmarssen, G. J. 1995. Evaluation of mass screening for cancer: a model-based approach [dissertation], Department of Public Health, Erasmus University Rotterdam, Rotterdam.

van Oortmarssen, G. J., J. D.F. Habbema, J.Th.N. Lubbe, and P. J. van der Maas. 1990a. A model-based analysis of the HIP project for breast cancer screening. Int J Cancer 46:207-213.

van Oortmarssen, G. J., J. D.F. Habbema, and M. van Ballegooijen. 1992. Predicting mortality from cervical cancer after negative smear test results [see comments]. BMJ 305:449-451.

van Oortmarssen, G. J., J. D.F. Habbema, P. J. van der Maas, H. J. de Koning, H. J. A. Collette, and et al. 1990b. A model for breast cancer screening. Cancer 66 (7):1601-1612.

Verbeek, A. L. M., R. Holland, F. Sturmans, J.H.C.L. Hendriks, M. Mravunac, and N. E. Day. 1984. Reduction of breast cancer mortality through mass screening with modern mammography. First results of the Nijmegen Project, 1975-1981. Lancet 1 (8388):1222-1224.

Vessey, M. 1994. Breast cancer screening 1991: Evidence and experience since the Forrest Report. Sheffield: Department of Health Advisory Committee, NHS Breast Screening Programme.

Warmerdam, P. G., H.J. de Koning, R. Boer, P. M. M. Beemsterboer, M. L. Dierks, and et al. 1997. Quantitative estimates of the impact of sensitivity and specificity in mammographic screening in Germany. J Epidemiol Community Health 51 (2):180-186.

Weinstein, M. C., J. E. Siegel, M. R. Gold, M. S. Kamlet, and L. B. Russell. 1996. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. JAMA 276 (15):1253-1258.

Weiss, N. S. 1983. Control definition in case-control studies of the efficacy of screening and diagnostic testing. Am J Epidemiol 118 (4):457-460.

Weiss, N. S., and D. Lazovich. 1996. Case-control studies of screening efficacy: the use of persons newly diagnosed with cancer who later sustain an unfavorable outcome. Am J Epidemiol 143 (4):319-322.

Weiss, N. S., B. McKnight, and N. G. Stevens. 1992. Approaches to the analysis of case-control studies of the efficacy of screening for cancer. Am J Epidemiol 135 (7):817-823.

Wesley, R., R. Sankaranarayanan, B. Mathew, B. Chandralekha, A. Aysha Beegum, and N. S. Amma. 1997. Evaluation of visual inspection as a screening test for cervical cancer. Br J Cancer 75 (3):436-440.

Whitmore, W. F., Jr., J. A. Warner, and I. M. Thompson, Jr. 1991. Expectant management of localized prostatic cancer. Cancer 67 (4):1091-1096.

Wildhagen, M. 1999. Costs and effects of genetic screening with application to cystic fibrosis and fragile X syndrome [dissertation], Department of Public Health, Erasmus University Rotterdam, Rotterdam.

ABOUT THE CHAPTERS

The following chapters have been published or are expected to be published elsewhere, some of these chapters have a final section 'additional discussion' that has not been published before.

2. de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. J Natl Cancer Inst 1995;87(16):1217-23.
3. Connor RJ, Boer R, Prorok PC, Weed DL. An investigation of design and bias issues in case-control studies of cancer screening using microsimulation. Am J Epidemiol. [in press]
4. Boer R, de Koning HJ, Beemsterboer PM, Warmerdam PG, Schroeder FH. A comparison of disease specific survival of patients who died of and who had newly diagnosed prostate cancer. J Urol 1997;157(5):1768-71.
5. Boer R, van Ballegooijen M, Ries L, Legler J, Feuer R, Habbema J. Ambiguities in calculating cancer patient survival: the seer experience for colorectal and prostate cancer. [submitted for publication]
6. Boer R, de Koning H, Threlfall A, Warmerdam P, Street A, Friedman E, et al. Cost effectiveness of shortening screening interval or extending age range of NHS breast screening programme: computer simulation study. BMJ 1998;317(7155):376-9.
7. Boer R, Marjolein van Ballegooijen, Willem-Jan Meerding, Veena Singh, Ashok Sehgal, J. Dik F. Habbema. Unaided visual examination versus Pap-smears for early detection of cervical cancer. [in abbreviated form in press in proceedings of Eurgin 2000 congress]
8. Boer R, de Koning HJ, van Oortmarssen GJ, van der Maas PJ. In search of the best upper age limit for breast cancer screening. Eur J Cancer 1995;31A(12):2040-3.
9. Boer R, de Koning HJ, van der Maas PJ. A longer breast carcinoma screening interval for women age older than 65 years? Cancer 1999;86(8):1506-10.
10. Boer R, de Koning H, van Ballegooijen M, van der Maas P. Important influences on effectiveness and costs to be considered in the evaluation of cancer screening. [in press in: Quantitative methods of evaluation of cancer screening]
11. Boer R, Warmerdam P, de Koning H, van Oortmarssen G. Extra incidence caused by mammographic screening [letter]. Lancet 1994;343(8903):979.
12. Boer R, de Koning H, van Oortmarssen G, Warmerdam P, van der Maas P. Stage distribution at first and repeat examinations in breast cancer screening. J Med Screen 1999;6(3):132-8.

ABOUT THE AUTHOR

Rob Boer was born 1960 in Vlaardingen, the Netherlands. After secondary school called atheneum-b at the Openbare Scholengemeenschap prof. Casimir from 1972 to 1978, also in Vlaardingen, he moved to Wageningen for his studies into plant pathology at the Agricultural University. This study was actually finished in 1984 but due to participation in the direction of the Wageningse Studenten Organisatie, the official end of his studies was not until 1985. From 1986 to 1987 he worked for the Foundation of Life Sciences and Society, then in Leiden, as educator/documentalist. From 1987 to 1988 he got educated as 'systems analyst' at SAVAG which led to work as programmer for the Bio-geographic Information Center of the Ministry of Agriculture, Nature Management, and Fisheries. In 1989 he started his work as informatician/statistician at the department of Public Health of Erasmus University Rotterdam where he is still working on the evaluation of breast, cervical, colorectal and prostate cancer screening.

THESES

1. Case-control studies can not provide sufficient evidence for the efficacy of early detection of cancer
2. Most women who undergo strong unfavourable health effects of screening for cancer will assume that they have benefited from that screening
3. Do not the same magnitudes appear larger to your sight when near, and smaller when at a distance? They will acknowledge that. And the same holds of thickness and number; also sounds, which are in themselves equal, are greater when near, and lesser when at a distance. They will grant that also. Now suppose happiness to consist in doing or choosing the greater, and in not doing or in avoiding the less, what would be the saving principle of human life? Would not the art of measuring be the saving principle; or would the power of appearance? (Protagoras, 380 BC, Plato)
4. Increasing the frequency of screening for breast cancer (within reasonable limits) results in a less favourable cost-effectiveness, but it improves the ratio of favourable and unfavourable health effects.
5. The decision to define the Dutch mass screening for breast cancer in women of age 70-75 years as an experiment is based on a misunderstanding and is therefore a violation of the law on proper medical treatment (art. 448, bk 7 van het Burgerlijk Wetboek)
6. A false balance is abomination to the LORD: but a just weight is his delight. (Proverbs 11:1)
7. Medical costs in life years gained by the intervention that is to be evaluated should not be included in the numerator of the cost-effectiveness ratio
8. A general custom of simulation is a vice (Of simulation and dissimulation, in: The essays, Francis Bacon, 1601)
9. More priority for identifying ineffective medical interventions will both improve public health and contribute to cost containment of medical care
10. The current preference for extra life years over better quality in present life years is appropriate
11. Resistance of plants against a particular pathogen is not durable if it results from genes of other species
12. The pressure to publish or perish can lead to undesirable delay in availability of research results
13. Spectacular results are more likely to arise from bad research than from good research
14. The presence of both safer sex promotion material as well as a cigarettes vending machine in the cafe of the COC-Rotterdam is based on wrong priorities
15. The term ‘big bang’ incorrectly suggests that there was something to hear at the time
16. The high profits made by scientific publishers are being paid by public money through scientific libraries
17. If men will permit themselves to think, as rational beings ought to think, nothing can appear more ridiculous and absurd, exclusive of all moral reflections, than to be at the expense of building navies, filling them with men, and then hauling them into the ocean, to try which can sink each other fastest. Peace, which costs nothing, is attended with infinitely more advantage, than any victory with all its expense. But this, though it best answers the purpose of nations, does not that of court governments, whose habited policy is pretence for taxation, places, and offices (The rights of man, Thomas Paine, 1792)
18. The world is flat