EVALUATION OF MASS SCREENING FOR CANCER
A MODEL-BASED APPROACH

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Oortmarssen, Gerrit Jan van

Evaluation of mass screening for cancer: a model-based approach.
Gerrit Jan van Oortmarssen. - [S.l.:s.n].
Proefschrift Rotterdam. -Met lit. opg. - Met samenvatting
in het Nederlands.
ISBN 90-9007840-1
Trefwoorden: borstkanker / baarmoederhalskanker / bevolkingsonderzoek.

Printed by Offsetdrukkerij Haveka, Alblasserdam, The Netherlands

EVALUATION OF MASS SCREENING FOR CANCER
A MODEL-BASED APPROACH

Evaluatie van bevolkingsonderzoek op kanker
een modelmatige benadering

PROEFSCHRIFT

TER VERGKRIJGING 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 VERDEDEGING ZAL PLAATSVINDEN OP
WOENSDAG 1 FEBRUARI 1995 OM 11:45 UUR

DOOR

Gerrit Jan van Oortmarssen

GEBOREN TE LOCHEM

PROMOTIECOMMISSIE

PROMOTOR:	Prof. dr. J.D.F. Habbema
OVERIGE LEDEN:	Prof. dr. D.E. Grobbee
	Prof. dr. R. Dekker
	Prof. dr. A.L.M. Verbeek

VOORWOORD

In de gezondheidszorg vindt modelbouw meer en meer ingang als hulpmiddel bij het nemen van beslissingen. Waarschijnlijk het best bekend zijn de besliskundige modellen die worden toegepast bij beslissingen rond diagnostiek en behandeling van individuele patienten. Deze worden echter relatief weinig gebruikt bij het nemen van praktische beslssingen. Ook op populatie-niveau worden, ter onderbouwing van strategische beleidsbeslissingen, modellen gebruikt voor het voorspellen van ontwikkelingen in morbiditeit, mortaliteit en gebruik van gezondheidszorg, en voor het analyseren van speciale programma's voor bestrijding van ziekten, zoals vaccinatie en bevolkingsonderzoek.

In dit proefschrift wordt de toepassing van modellen voor analyse van bevolkingsonderzoek dat zich richt op vroege opsporing van borstkanker en van baarmoederhalskanker.

De engelstalige hoofdtekst en samenvatting geeft een wetenschappelijke verantwoording van mijn onderzoekswerk op dit terrein. In de Nederlandse samenvatting heb ik getracht om in voor iedereen te begrijpen bewoordingen de inhoud en het belang van het onderzoek onder woorden te brengen.

PREFACE

Models are gradually becoming an accepted tool in decision making in health care. Probably best known are the decision-tree like models that aim at assisting a physician in selecting appropriate diagnostic and treatment procedures for individual patients. At population level, epidemiometric models are being used to predict future trends in morbidity, mortality and demand for health care facilities, and to evaluate disease control programmes.

The beginning of my involvement in epidemiometric modelling at the Department of Public Health and Social Medicine was in 1977, when the SCRMOD computer simulation program developed in the U.K. by E.G. Knox was used to predict the effects of cervical cancer screening in the Netherlands. The results were presented at the symposium organized at the event of the retirement of professor Burema. Based on this experience, it was decided to start developing our own computer simulation programme, which was given the acronym MISCAN (MIcrosimulation SCreening ANalysis). This was the start of the research project "Decision making about mass screening for cancer".

MISCAN-based models for breast and cervical cancer screening have been used for analyzing some of the major screening projects in the world and in The Netherlands: the HIP randomized controlled trial of breast cancer screening, the cervical cancer screening programme in British Columbia, the Dutch pilot study for cervical cancer screening, the breast cancer screening projects in Nijmegen and Utrecht, and the Swedish randomized trials for breast cancer screening. These analyses yielded models that subsequently have been used in comprehensive cost-effectiveness evaluation studies of cervical cancer screening and breast cancer screening, first for The Netherlands but more recently also for other countries.

The practical relevance of our modelling work for cancer screening has become evident after we reported the results of the cost-effectiveness studies. The outcomes for breast cancer screening were in support for the screening strategy as proposed in The Netherlands. Both our favourable conclusions about the ratio of costs and effects, and the finding that the adverse effects of a well-organized and conducted programme would be relatively minor in comparison with the anticipated mortality reduction, have contributed to the decision to start nationwide screening. The results of the evaluation of cervical cancer screening challenged the current practice of screening in The Netherlands, and have clearly influenced the decision for a change in screening policy.

Publication of our cost-effectiveness results has also resulted in a demand for this type of analyses in other countries. Collaborative efforts have resulted in cost-effectiveness analyses of breast cancer screening in Australia and Germany, and at present similar analyses are carried out in cooperation with groups in Italy and Spain.

The experience gathered in modelling screening for cancer has been used in a number of modelling efforts in which I have been involved: models for tropical disease control and for public health problems in developed countries.

A direct descendant of MISCAN is the HUMPAR model for evaluation of control programmes for the tropical infectious disease onchocerciasis (river blindness). The model has been used in analyzing and predicting the decreasing trend in infection load as a result of large scale vector control carried out in the Onchocerciasis Control Programme (OCP) in Western Africa. The relatively simple HUMPAR model has now been superseded by the comprehensive ONCHOSIM model, which describes the full transmission cycle of river blindness, the development of blindness, and the impact of different intervention strategies based on vector control or chemotherapy (ivermectin). The ONCHOSIM model is not only highly complex, it has also been quite influential, since it has been used extensively in preparing policy decisions about control activities within the OCP. Similar models are under development for other tropical infectious diseases: lymphatic filariasis, schistosomiasis, and leprosy.

Examples of Public Health models oriented to the problems in the Netherlands include the PREVENT and TAM models, and a model for predicting trends in accident mortality. The PREVENT describes the interaction between risk factors and diseases in a population, and can be used to predict the health effects of preventive measures. The TAM model can be viewed as a ambitious extension of PREVENT, and aims at giving a general description of the public health and the demand for health care in The Netherlands.

This thesis describes the first, essential step in model-based evaluation of screening for cancer: the construction, estimation and validation of epidemiometric models for breast cancer screening and cervical cancer screening. A series of articles on evaluation of screening is preceded by a general overview of modelling and its application to breast cancer screening and cervical cancer screening (Chapter I). Chapters II-VI provide examples of analyzing results of breast cancer and cervical cancer screening projects. In the concluding Chapter (VII), the results of the model-based analysis and directions for further investigation are discussed.

I. INTRODUCTION

Evaluation of screening for cancer

The main goal in evaluation of screening for cancer is to assist in decision making about a screening program: Should it be initiated at all? What screening policies can be recommended: what age groups, what frequency of screening. Should special attention be paid to high risk groups? If a screening program is already running, should screening be continued in view of the results? Should the present policy be changed?

In this chapter, I will describe the complexities involved in answering these questions. These difficulties lead to the conclusion that models are indispensable in the interpretation of observed results of screening and in the prediction of effects and costs of different screening policies.

Screening for cancer

From a public health perspective, the most attractive method of cancer control is primary prevention by reducing the exposure of the population to risk factors. Primary prevention will reduce both mortality and morbidity. Secondary prevention by means of early detection and treatment is a method of cancer control which aims at a substantial reduction in mortality from a specific type of cancer. A secondary aim, which will be more difficult to achieve, is to reduce morbidity. Mass screening implies examination of a population of apparently healthy individuals, using a suitable screening test to identify those who are likely to have an early stage of cancer. Mass screening programmes for cervical cancer and breast cancer exist in many countries, especially in Northern and Western Europe.

Evaluation

A inherent problem of mass screening is that adverse effects will occur such as false positive test results or treatment of screen-detected cancers which would never have been diagnosed in the absence of screening. These will strike persons that will not have any of the benefits of screening, such as less intensive treatment, prevention of invasive cancer and/or of death from the cancer. The existence of a group of persons that will suffer from being screened, dictates that screening programmes should be evaluated carefully to assess whether the anticipated favourable effects outweigh these adverse effects. Another reason for careful evaluation is important in an era of cost containment: mass screening programmes are expensive, and although considerable financial savings in treatment costs are possible, they will in general be much smaller than the extra costs of screening. Furthermore, whereas most of the costs (and the adverse effects) are made when performing screening tests and in follow-up of screen-positives, the possible savings will only be acquired years later.

Assessment of the effects of screening is far from straightforward (Morrison, 1992; Habbema *et al.*, 1982ab). The main effects, such as prevented deaths on the one hand, and unnecessary treated screen-detected cases on the other hand, cannot be established directly on basis of data recorded in screening registries (positive screening results, screen-detected cancers) and cancer registries (clinically detected cancers, survival). Only a careful comparison of a screened population and an unscreened population, by means of a randomized controlled community trial may give a decisive answer on the question of mortality reduction.

The difficulties involved in decision making about mass screening become evident when comparing the (recommended) screening policies for cervical cancer and for breast cancer in different countries, see Tables I.1 and I.2. The discussion regarding breast cancer screening policies (Table I.1) is focusing on the issue whether women younger than 50 should be screened. The variation between policies is largely due to different interpretation of the outcomes of randomized trials and some non-randomized pilot studies.

Table I.1 Recommended and actual screening policies for breast cancer.

Country	Age-range	Interval (years)	Source
The Netherlands	50-68	2	Van der Maas *et al.*(1989)
USA	40+	<50:1 or 2, >50:1	NCI/ACS (1987)
United Kingdom	50-64	3	Vessey (1991)
Finland	50-59(-69)	2	Hakama (1988), Pamilo *et al.* (1993)

The large differences in recommenced cervical cancer screening policies (Table 1.2) cannot be explained easily. One reason of variation might be differences in the objectives of screening. Focusing on the (intermediate) objective of maximizing the number of screen-detected preinvasive stages would lead to emphasis on screening at young ages, and with relatively short intervals.

Minimizing the incidence of invasive cervical cancer and subsequent mortality would result in a shift to higher ages and longer intervals.

A further complication in many countries is that the actual screening pattern differs substantially from the recommended schedule, with a considerable proportion of smears being made in very young women and/or at very short intervals. However, it seems that at least in Europe, a consensus is growing about the age range (from ±25 to ±65) and the screening interval (at least 3 years), see Europe Against Cancer (1993).

Table I.2 Differences in screening policies for cervical cancer

Country	Age-range	Interval (years)	Source
The Netherlands, old	35-53	3	Evaluation Committee (1990)
The Netherlands, new	30-60	5	
USA	18+	1	NCI (1987)
United Kingdom	20-65	5	Holland and Stewart (1990)
Sweden	30-49	4	Läärä *et al.* (1987)
Finland	30-55	5	
Iceland	25-69	2-3	
Canada	18-60	<35:1, >35:5	Canadian Task Force (1982)
European Union	25-65	3-5	Europe Against Cancer (1993)

Retrospective and prospective evaluation

In evaluation of mass screening for cancer, a distinction can be made between retrospective evaluation of existing screening projects, and prospective evaluation of proposed screening policies. The goal of retrospective evaluation is to analyze data from screening projects in order to explain the observed screening results, and (if possible) to draw conclusions about the effects of screening. The findings can subsequently be used in prospective evaluation, which aims at making predictions about the effects of screening policies. This may lead to recommendations about starting or continuation of a screening programme and about efficient screening policies. Evaluation should continue when the mass screening is an established component of health care ("monitoring"), in order to check whether the expected benefits have indeed been accomplished.

Analysis of screening results

A prerequisite for a detailed retrospective evaluation (or analysis) is that empirical data from well-designed experimental studies are available. Ideally, randomized controlled trials (RCT's) should have been conducted. In the absence of a RCT, or in addition to it, longitudinal data from screening programmes, including cancer incidence and mortality are useful. An indication for the effect of screening can also be obtained from case-control studies.

Retrospective evaluation first of all requires a careful statistical analysis of the available data. In addition, model-based analyses may be carried out to make inferences about the parameters governing the results and effects of screening. The most important parameters describe the so-called *underlying processes* (Bross and Blumenson, 1968; Eddy and Shwartz, 1982): the preclinical course of the disease, the quality of the screening test, the impact of early detection and treatment on a patients' prognosis, and the association between disease risk and participation in screening. These processes cannot be observed directly. Statistical comparison of data from screened and unscreened population may already give an indication of regression (Boyes *et al.*, 1982), the duration of preclinical stages (Hakama *et al.*, 1986a; Day, 1985a) and the sensitivity of the screening test (Day, 1985b).

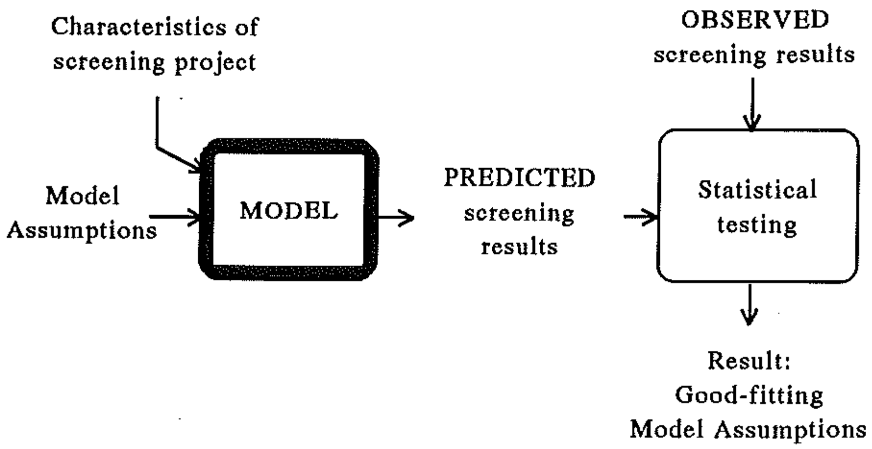
Characteristics of screening project
Model Assumptions
MODEL
PREDICTED screening results
OBSERVED screening results
Statistical testing
Result: Good-fitting Model Assumptions

Figure Ia Model-based analysis of mass screening results.

Modelling can improve these direct methods in two ways. First, instead of estimating the parameters of these underlying processes separately, joint estimates (including their correlation) are derived, for example regarding the duration of the screen-detectable stages and the sensitivity of the screening test. Second, more data can be included in the analysis, e.g. detection rates at successive screenings, interval cancer rates, clinical incidence in unscreened persons.

A general procedure for model-based analysis is outlined in Figure Ia. Different model assumptions are investigated by means of a statistical comparison of the screening results predicted by the model and the observed results of screening projects. The choice of the model structure and the selection of observed screening results are of critical importance for the quality of the model. It is hoped that the result of this modelling stage will be a set of model-specifications that are in good agreement with available empirical screening data. Subsequently, these good-fitting model assumptions can be utilized in prospective evaluation of screening policies, see Figure Ib.

Prospective evaluation

In view of the many parameters that determine the direct results and long-term effects of screening for cancer, modelling is the method of choice for evaluating the impact of proposed screening policies. In prospective evaluation, policies will be compared on basis of the expected favourable and unfavourable effects, and their costs, in a specific population, see Figure Ib. The model predictions can be used to produce a complete overview of the expected benefits and potential risks, which may serve as a guidance for the individual person who is invited for screening. In addition, predictions may be made regarding the impact of screening on existing health care services.

In this thesis I will focus on the first stage of modelling as depicted in Figure Ia: development and testing of models. I will describe two different approaches for model development and parameter estimation: a statistical-numerical approach, and a simulation approach which uses the MISCAN programme. These two methods represent different trade-offs between methodological correctness on the one hand and medical/epidemiological plausibility ("face-validity") of the model assumptions on the other. Ideally, these two approaches are combined, starting with a statistical-numerical analysis which is then followed by a more comprehensive simulation modelling effort (Chapter IV of this thesis; Van Ballegooijen *et al.*, 1993).

Application of these models to the second stage (Figure Ib) of cost-effectiveness evaluation requires extensions in which the cost of the screening

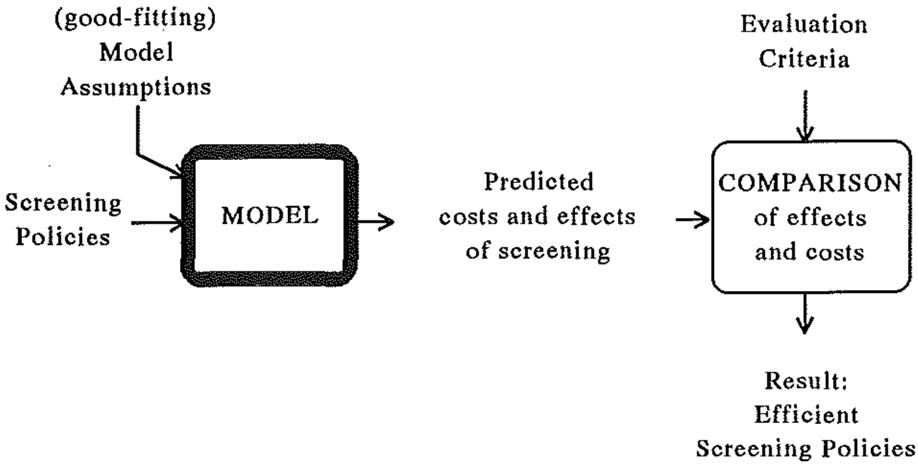

Figure Ib Prospective evaluation of screening policies.

organisation, the medical procedures, and the quality of life in different phases are added to the model. Modelling of these extensions is not a subject of this thesis, the assumptions made and the resulting predictions have been described elsewhere (Van der Maas *et al.*, 1989; Van Ballegooijen *et al.*, 1990,1992a; De Koning *et al.*, 1990a,1991; Koopmanschap *et al.*, 1990a).

An important aspect in the development of a model is the implementation in computer programme(s). Although much effort has been devoted to the development of the MISCAN computer programme (Van Oortmarssen *et al.*, 1982; Habbema *et al.*, 1984,1987), only a brief overview of the program and the simulation procedure will be presented in this thesis.

Analysis of screening projects

In analyzing data from screening activities for deriving estimates of the effect of screening, a distinction has to be made between experimental studies, i.e. randomized controlled trials, and non-experimental studies, for example cohort studies, case-control studies and geographical correlation. Each study design requires special methods of analysis. Designs will also differ with respect to the degree in which systematic bias in estimates can be avoided (Morrison, 1992).

Biases in effect estimates are best illustrated for a method of evaluating the effect of screening which used to be quite common in the early days of screening: comparison of survival. Screen-detected and clinically diagnosed cases are compared on basis of data recorded in survival registries. Survival outcomes, for example the proportion of persons that survives after 5 years, or the full distribution of time between diagnosis and death, are inadequate for analysis of the effect of screening because of the disturbing influence of several sources of bias: *lead time* bias, *length* bias, *patient self selection* bias, and *overdiagnosis* bias. Even if screening would have no effect at all, the survival would be more favourable for screen-detected cases than for clinically diagnosed cancers. The survival time, i.e. the time interval between diagnosis and death, will be longer in screen-detected cases, because of the addition of the *lead-time* period between detection by screening and the moment at which the cancer would have been diagnosed clinically in the absence of screening. In addition, screening will tend to pick up relatively slow developing cancers, and miss fast growing ones: *length biased sampling.* And it could well be that slowly growing cancers have a better than average survival, which leads to further bias in the survival time comparison.

Patient self selection occurs when health-conscious persons who are more likely to participate in screening, also have lower mortality risks in the absence of screening. This is not as unlikely as it may seem. People who are alert to slight symptoms may be more inclined to participate in a screening programme than persons who postpone medical examination in spite of alarming symptoms. This will again lead to a seemingly longer survival time for screen-detected cases.

Overdiagnosis bias of survival time will occur when cases are detected by screening which would never have been diagnosed in these absence of screening because the condition ("pseudodisease") will only be detected by a screening test, but never lead to severe symptoms. If these cancers have a good prognosis, the lethality and cancer mortality of screen-detected cancers will be lower than in an unscreened population, even if screening has no effect at all. Another type of overdiagnosis of screen-detected cancers, which is especially important at older ages, occurs in case of death from other causes during the lead time period. But this does not necessarily lead to biased survival outcomes, since it will not affect relative survival figures. The negative health effects of "unnecessary" diagnosis and treatment of these cancers are of course important in evaluation of screening at older ages (Boer *et al.*, 1994b).

The impact of these four sources of biases of the survival can be large (Habbema *et al.*, 1983; Walter and Stitt, 1987; Shwartz, 1980; Morrison, 1992; Pelikan and Moskowitz, 1993). Thus, it is impossible to use survival time comparisons for estimating the true effect of screening because of biases that are specific for screening. Other, more laborious methods should be used that are less vulnerable for bias, although some amount of bias will always will be present, either the screening-related biases or general problems in measuring the effect of interventions in health care, like for example placebo effects.

Randomized controlled trials

A randomized controlled design is the method of choice for obtaining objective estimates of the effect of screening in reducing mortality. Still, bias cannot be excluded because of practical constraints in designing a RCT. Ideally, one would like to use total mortality as an endpoint to avoid bias because of possible side-effects of screening, see for example the higher mortality from suicide in the Kaiser Health Plan trial (Friedman *et al.*, 1986), and also because of differences in classification of mortality. But using total mortality requires a much larger study population than the already giant population needed for demonstrating a significant reduction in mortality from the cancer under consideration. Other practical problems concern the level of randomization (individual, general practitioner, county), "contamination" with screening in the control group, and the cessation of the trial (stop screening in the screening group or introduce screening in the control group) with the associated problem of avoiding extra cases in the study group, a special kind of overdiagnosis bias.

A delayed entry RCT uses a slightly different approach in that screening will be offered to the total study population, but the screening starts later in the control group than in the study group. The delayed entry design is also useful in a non-experimental screening programme. Introducing a (nationwide) mass screening requires a build-up period, and the resulting differences in time-at-entry into the programme can be used in analyzing the results and effects on incidence and mortality. This approach is currently tested in the Finnish breast cancer screening programme. First all women born in even years have been invited, and the women in the adjacent odd-years serve as control population during the first part of the build-up period (Hakama, 1988; Hakama *et al.*, 1991).

Cohort studies: comprehensive analysis of screening data

Comprehensive analysis of screening data from non-randomized screening programmes may provide a direct estimate of the effect on mortality, but may also be used for estimating other key factors in the evaluation of screening. In comparison with RCT's, the resulting estimates may be severely biased, especially because of the self-selection bias: participants in screening may have a lower risk of having the disease than non-participants, and since participating women tend to be relative health-conscious, they might also have a better prognosis in the absence of screening. The best examples of this approach address the evaluation of cervical cancer screening, which is no surprise given the absence of RCT's for this type of cancer screening.

The British Columbia Cohort Study (Boyes *et al.*, 1982) was aimed at deriving estimates for the incidence of preclinical stages of cervical cancer, and assessing the proportion of these lesions that regress spontaneously. Two birth cohorts were selected, and for women in these cohorts screening records and cervical cancer records were linked for a 20-year period (1950-1969). This enabled estimation of the age-specific incidence of preinvasive cancer. By comparing this incidence with the pre-screening, cross-sectional incidence of invasive cancer, the proportion of regressive cases could be estimated. A model-based analysis of the same data is reported in Chapter IV.

The pilot projects for breast cancer screening and cervical cancer screening in the Netherlands have also been analyzed extensively (Peeters *et al.*, 1989; Peer *et al.*, 1994; De Waard, 1984; Collette *et al.*, 1992ab; Evaluation Committee, 1989; Van der Graaf, 1987). In Chapter III a model for breast cancer screening is presented which is based on the data from the projects in Utrecht (DOM project) and Nijmegen.

Correlation studies

Correlation studies for the effect of screening compare regions or countries, attempting to correlate screening activities with observed changes in clinical incidence and mortality. Convincing evidence for the effectiveness of screening for cervical cancer has been obtained in this way by comparing data from the Nordic countries (Läärä *et al.*, 1987), and by comparing provinces of Canada (Miller *et al.*, 1976).

Case-control studies

In case-control studies, the exposure to screening of all cancer *cases* in a well defined population and period is compared with the exposure of non-cancer *controls*. Usually, the controls are matched against the cases on bases of known or suspected risk determinants (e.g. age, neighbourhood). The screening experience of cases and controls is compared, and the odds ratio is interpreted as an indicator for the effectiveness of screening in reducing the number of cases. This offers a relative quick method for evaluation of screening projects that do not use a randomized design.

A specific problem is the definition of *cases*. The definition should reflect the condition which is to be prevented, and *not* detected, by screening, i.e., death from cancer or an advanced stage of cancer that will be diagnosed anyhow. However, this may in practice lead to difficulties in collecting the data regarding the screening history. Use of alternative case definitions, e.g. all diagnosed cancers, will give rise to biased estimates.

Case-control estimates for the effectiveness of screening are also likely to be biased because of self-selection bias. Another potentially important source of bias which especially applies to case-control studies of screening is the *healthy screenee bias.* Again consider the situation that early detection does not affect risk of death from the cancer. Then, cases will on average have a smaller number of screenings before the moment of death than comparable (matched) controls, because of the period of disease which precedes death, and in which persons are not eligible for screening. Comparison of crude exposure to screening would suggest a beneficial effect. More appropriate measures of exposure have been proposed and used. The adaptations needed depend on the characteristics of screening patterns in a population, and the kind of relation considered, e.g. ever versus never screened, or a dose-response relation.

See (Morrison, 1982,1992; Brookmeyer *et al.*; 1986; Sasco *et al.*, 1986; Friedman and Dubin, 1991; Weiss, 1994) for a discussion of these and other problems involved in applying the case-control design to cancer screening.

Many case-control studies have been reported, both for breast cancer screening, in which mortality usually is the end-point (e.g., Collette *et al.*, 1984,1992; Verbeek *et al.*, 1984; Palli *et al.*, 1986), and for cervical cancer screening, in which invasive cancer is usually taken as end-point (e.g., Clarke and Anderson, 1979; Macgregor *et al.*, 1985; Lynge and Poll, 1986; Van der Graaf, 1987). In Chapters V and VI it will be shown that the use of invasive cervical cancer as case definition will lead to biased estimates of the long-term impact of cervical cancer screening on mortality.

Meta analysis

Given the absence of a randomized controlled trial of cervical cancer screening, a IARC/WHO study group concluded that a combined analysis of major non-randomized screening studies would give convincing estimates of the reduction in risk of invasive cervical cancer after one or more negative Pap smears (Hakama *et al.*, 1986a). The advantage of this approach can be seen by inspecting the estimates for the separate projects, which are less convincing because of the relative small numbers of cancers in each of the projects. We have used the model constructed on basis of the British Columbia cohort study to analyze the outcomes of the IARC study, and demonstrated that a single quantification of the underlying processes gives a good explanation of both data sets.

For breast cancer screening, a number of combined analyses of the outcomes of the randomized trials have been reported (Rutqvist *et al.*, 1990; Glasziou, 1992; Vessey, 1991; Fletcher *et al.*, 1993; Screening brief, 1994). However, a weak point of these analyses is that the obvious differences in the characteristics of the trials are not (or insufficiently) taken into account. This limitation can only be overcome by using explicit models.

Surrogate outcome measures

The enormous size and long duration of randomized controlled trials of cancer size prohibits testing of a range of policy options, and also means that major changes occurred with respect to the screening test, diagnostic tests, or treatment possibilities at the time when the results become available. This has given rise to search for alternative methods for empirically testing policy options. One relative recent method which is to consider *surrogate (or intermediate) outcome measures* instead of the real outcome measure (mortality) (Day, 1991; Morrison, 1992). First, the relation between the surrogate outcome as determinant of the real outcome has to be established. Policy options are then compared in studies which take the surrogate outcome as endpoint, and the relation with the real endpoint is used to translate the intermediate outcome into the final effect.

Model based analysis

In the studies mentioned in the preceding section, model-based reasoning is used in evaluation of screening projects, for example in searching for appropriate measures and formulas for analyzing screening data. But the results are calculated without explicit use of models. The outcomes of the studies are mostly in terms of risk reduction and therefore have a direct interpretation both for individuals who contemplate screening and in assessing the public health impact of screening. However, the outcomes are specific for the characteristics of the projects under study and cannot be generalized easily to other circumstances regarding screening policy, epidemiological situation (e.g., clinical stage distribution), etc.

The goal of model-based analysis of screening results is to test hypotheses and to make inferences about the underlying dynamic processes which cannot be observed directly, but may be assumed to apply to other circumstances as well. Two underlying processes that together relate the screening policy to the effect of screening should be distinguished:

(1) DETECTION, the process of detecting a cancer in an early stage;

(2) PROGNOSIS, the improvement in prognosis resulting from early detection.

These two processes are the building blocks of the basic cancer screening model structure which comprises 9 states, see Figure Ic and Table I.3. Persons start in state NO CANCER, and the large majority will not have a cancer history and experience only one transition to state DEATH FROM OTHER CAUSES. The two preclinical screen-detectable states at the left-hand side will normally remain unnoticed in the absence of screening, and the cancer will be diagnosed after presenting symptoms (state CLINICAL INVASIVE). Treatment may result in CURE or can be unsuccessful and the person dies from the cancer (DEATH FROM CANCER). Screening may result in early detection of stages SCREEN-DETECTABLE NONINVASIVE and SCREEN-DETECTABLE INVASIVE. In the figure, it is assumed that treatment of SCREEN-DETECTED NON-INVASIVE cancers will always result in cure. Persons with treated SCREEN-DETECTED INVASIVE cancers will have a lower risk of dying from the cancer than treated CLINICAL INVASIVE cancer cases.

The basic events that would be recorded in a cancer registry are indicated by E_1-E_5. The ages and results of the screening tests (G_1...G_k) would be available from a screening registry. The corresponding detection rates at successive screening examinations, the incidence of interval cancers, and the clinical incidence in unscreened persons can be used to estimate the "deep" parameters of the underlying processes. The parameters of the DETECTION part characterize transition rates in the preclinical disease process (p_1-p_4) and the sensitivity of the screening test (p_5, p_6). The remaining parameter (p_7) of the PROGNOSIS part describes the relation between earliness of detection and mortality risk, which can be estimated from observed mortality difference between screened and unscreened populations, ideally from the mortality rates in the study- and control group of a randomized controlled trial.

The surface parameters D_1-D_4 can usually be based directly on local data from the area under consideration. Indeed, these parameters may differ considerably between areas: the false positive rate D_2 in breast cancer screening in The Netherlands and Sweden is much lower than in countries like the USA, and also survival rates D_3 show significant variation between countries. Some of the deep parameters are likely to have a local component as well. For example the level and age-distribution of the onset depends on risk factors in the population, and the test sensitivity may be influenced by local factors, e.g. training, quality control, and follow-up protocols. Both in transferring a model to other areas and in making predictions one should take possible changes in parameter quantifications into consideration.

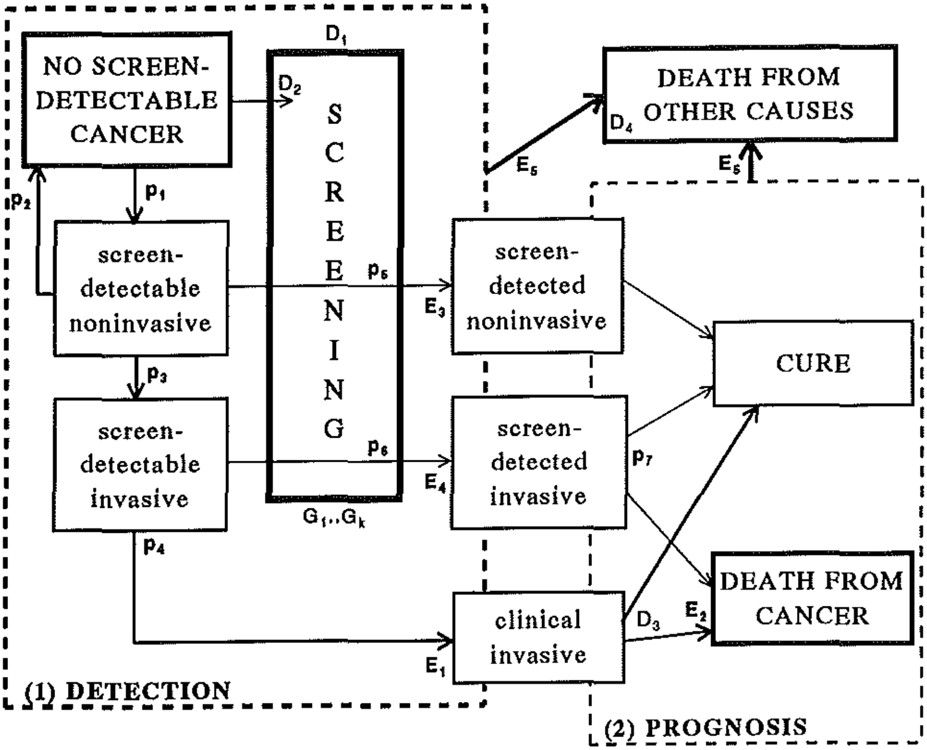

Figure Ic Basic structure of a model for cancer screening.

Table I.3 Parameters and events in the basic model of cancer screening.

Deep parameters		Surface parameters		Events	
p_1	Onset	D_1	Participation	E_1	Clinical Incidence
p_2	Regression of pre-invasive stages	D_2	False Positives	E_2	Mortality from cancer
p_3	Progression of pre-invasive stages	D_3	Clinical Survival	E_3	Detection Rate Pre-Invasive
p_4	Duration of preclinical invasive stage	D_4	Death from other causes	E_4	Detection Rate Invasive
p_5	Test Sensitivity, pre-invasive stage			E_5	Death from other causes
p_6	Test Sensitivity, invasive stage	SCREENING HISTORY			
p_7	Improvement in prognosis	$G_1...G_k$		Ages and results of k screening tests	

For specific applications, this basic model will should be refined to reflect the characteristics of both the cancer under study and the screening projects that are analyzed. For cervical cancer screening the primary aim is detection of the non-invasive stages of the disease, therefore the non-invasive state of the model can be subdivided in dysplasia and CIS (or, CIN I/II/III). For breast cancer, the screen-detectable invasive state is the most important. By subdividing this state according to tumour diameter and/or lymph node status, the shift in stage distribution resulting from screening according to the model can be compared with observed stage distributions in screen-detected and interval cancers. And the assumptions about improvement in prognosis can be checked against stage-specific survival data, both for screen-detected cancer and clinical cancers.

Modelling approaches

A large number of cancer screening models have been published and reviewed (Eddy and Shwartz, 1982; Prorok, 1986,1988; Alexander, 1989; Baker *et al.*, 1991).

The models differ in objective, in comprehensiveness, and in (numerical) methods used in implementation. Objectives are: (a) formal (mathematical) description of the processes involved; (b) parameter estimation and hypothesis testing; (c) prediction of the results of screening policies; (d) cost-effectiveness analysis; (e) optimization of policies. In objectives (a-c), the goal may be to study one or more specific aspects, which leads to a partial model, for example models that only address the DETECTION part of figure I.c. Still, these partial models may be very detailed. Comprehensive models are required for cost-effectiveness analyses and for optimization of screening strategies.

The objectives (a)..(e) might be considered as a sequence in which models development starts from a largely theoretical basis, followed by estimation and validation, and can then be applied as a decision support tool in public health decision making. In practice this trajectory has never been fully completed, although the modelling efforts of Eddy (1980,1981,1987,1988,1990), Shwartz (1978ab,1980,1981,1992), Gustafsson and Adami (1989,1990,1992) and our MISCAN work address most of the objectives.

Mathematical descriptions of screening for cancer resulted in highly theoretical work (Blumenson, 1977; Louis *et al.*, 1978; Prorok, 1976ab), but also served as a starting point for analysis of screening data (Day and Walter, 1983; Gustafsson, 1986) and prediction of screening policies (Eddy, 1980; Shwartz, 1978b) and thus influenced applications of great practical relevance (Walter and Day, 1983;

Eddy, 1980,1981,1988,1990; Gustafsson and Adami, 1989,1990,1992; Shwartz, 1978a,1980,1992).

Two classes of models only aim at certain objectives: descriptive statistical models and macrosimulation models.

A considerable number of descriptive statistical models has been developed for addressing specific questions regarding components of the general structure in Figure Ic. Examples are the ROC models describing the change in false-positive and false-negative rates when different criteria are used in classifying results of screening tests, and the Age-Period-Cohort model for describing temporal changes in incidence and mortality.

Macro-simulation (or discrete-time Markov) models are a type of dynamic stochastic models that are mainly used for predictions of the results of screening policies. "Stochastic" implies that transitions between disease states are described by probability distributions. The term "dynamic" is used to stress the importance of time-related events and their associations, e.g. aging of persons, development of disease including dwelling time of preclinical screen-detectable stages, survival time, and timing of screening examinations. The best example is the model of Knox (1973,1975) which has also been used by our group (Habbema *et al.*, 1978) and in other countries (Yu *et al.*, 1982; Stevenson *et al.*, 1991; Carter *et al.*, 1993). In macro-simulation the complete distribution over all possible combinations of states (including disease states, dwelling time in state, screening history, birth cohort, and age) is calculated at each time-step (usually 1 year). The number of combinations will rapidly become very high (for example 80 ages x 25 dwelling times x 25 disease states x 16 types of screening histories = 800,000 combinations), and require much computer memory and computer time when these calculations have to be made for e.g. 80 yearly birth cohorts. Also, the 1 year time-step may be too crude for some output, e.g. of interval cancers. For these reasons, macrosimulation is not very well suited for analysis of screening data.

Two types of numerical modelling approaches are more useful for comprehensive analysis of screening data: statistical-numerical models and microsimulation models. Essentially, they differ only in the numerical technique which is applied to the dynamic stochastic cancer screening model shown in fig Ic. The most elementary dynamic model, involving 1 disease state and 1 screening test, was described mathematically by Zelen and Feinleib (1969), and demonstrates the existence of length biased sampling, the way it influences the lead time, and the relation with the distribution of the preclinical screen-detectable state.

More refined dynamic models will rapidly become too complex for analytical solution. Then, *statistical-numerical* techniques can be applied for obtaining estimated values for the parameters of the model. The numerical approach used is comparable to macrosimulation models, with two major differences: the time-perspective and the time-step. Macro-simulation proceeds in a forward direction. Starting from an initial distribution over all states, this complete distribution is updated at successive time steps. This time step is fixed (usually 1 year), which might lead to inaccuracies with respect to the possibility of multiple events occurring during a single step. In the statistical-numerical approach a backward direction is adopted. For each outcome variable of the model that has to be calculated, equations are derived that integrate over all possible earlier states. Numerical integration routines are used which can achieve high accuracy. This approach is more efficient in case of limited number of (aggregated) observed data points that have to be fitted by the model, and for relatively simple models.

Most statistical-numerical models are confined to the DETECTION box in Figure Ic. The best known example is the one-stage/multi screening model of Day and Walter (Day and Walter, 1984; Day *et al.*, 1988; Walter and Day, 1983) which especially applies to breast cancer screening (or in general, to screening for a disease with a relatively short, progressive preclinical stage). The parameters (mean duration p_4 and sensitivity p_6) are estimated from detection rates (E_4) and interval cancers (E_I) in subsequent screening rounds. Application to data from different screening projects showed that the duration of the preclinical stage of breast cancer can be adequately described by an exponential distribution, implying considerable variability, i.e., both high proportions of fast growing and of very slowly progressing cancers.

This model has been generalized in different ways: two preclinical states (Brookmeyer and Day, 1987), time-dependent sensitivity (Verbeek and Straatman, 1988), multi-state/two screening methods (Alexander, 1989a). Our statistical/numerical cervical cancer model (described in Chapters IV-VI) may also be regarded as an extension of the model of Day and Walter. It consists of the complete DETECTION box in Figure Ic with its 6 parameters $p_1..p_6$, and has 4 additional parameters: age-dependent pre-clinical incidence, age-dependent regression, variability of the preinvasive stage, and risk difference regarding onset of pre-invasive stages between participants and non-participant in screening. A similar cervical cancer screening model has been used to analyze Swedish screening, incidence, and mortality data (Gustafsson and Adami, 1989,1990,1992). This model includes the PROGNOSIS box of Figure Ic. Another model which includes both the DETECTION and the PROGNOSIS boxes has been developed for lung cancer screening (Flehinger and Kimmel, 1987,1993). Partial breast cancer screening models addressing the PROGNOSIS part have been proposed by Habbema *et al.* (1983), Walter & Stitt (1987) and Connor *et al.*(1989).

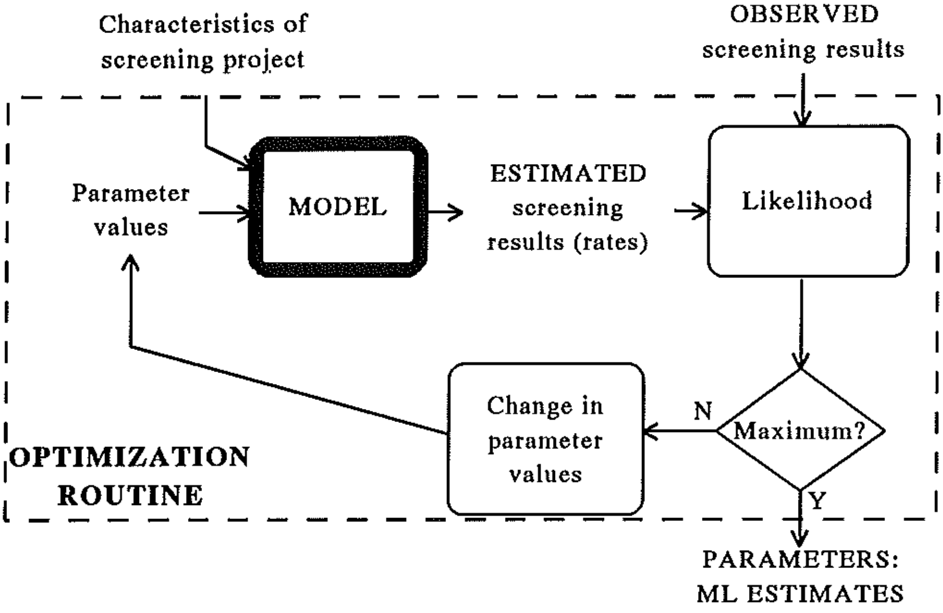
Characteristics of screening project
OBSERVED screening results
Parameter values
MODEL
ESTIMATED screening results (rates)
Likelihood
Change in parameter values
N
Maximum?
Y
OPTIMIZATION ROUTINE
PARAMETERS: ML ESTIMATES

Figure Id Statistical-numerical model for analyzing screening data. Parameters are estimated in an iterative procedure involving calculation of the model outcomes and of the likelihood, and subsequent adaptation of parameter values.

In the statistical-numerical approach, the elements of the general analysis scheme (Figure Ib) are combined in an iterative procedure which is shown in Figure Id. The MODEL step requires writing down the equations which express the expected rates for the screening data as a function of the parameter values (see for example Chapter IV, appendix B). In general, these equations will be too complex to be solved analytically, and numerical approximation methods are used to calculate expected rates from the model for given parameter values. Parameter estimation is based on maximization of the likelihood, assuming that the observed numbers of cases are realisations of a Poisson distribution with mean value = predicted rate x observed number of persons "at risk". Starting with a set of initial parameter values, the iterative optimization routine will result in maximum likelihood (ML) estimates of the parameters. Hypotheses can be tested by means of the likelihood-ratio test. The goodness of fit is checked by looking at the deviance and the Pearson Chi-square.

The statistical-numerical method is relatively fast in terms of computer time. However, it requires substantial effort for deriving the formulas involved, which will rapidly become complicated when a model is extended. This imposes restrictions to the complexity of the model, and also hampers testing of the robustness of the model assumptions. The main advantage of the approach is that standard maximum-likelihood methods can be used for testing hypotheses and deriving confidence regions.

A completely different approach is used in *microsimulation*, a technique based on generating large numbers of random trajectories, governed by the probability distributions which describe the processes in the model. Applied to cancer screening, this means that individual life-histories of (fictitious) persons are generated, including a possible disease history and the results and effects of screening examinations. A simulated life-history is characterized by the events which are related to cancer screening (see G_1.. and E_1.. in fig Ic). A file of these life-histories can be considered as simulated population-, cancer-, and screening registry, and model output is calculated correspondingly by counting the number of cases and the number of person(year)s at risk.

The output of a microsimulation run will be subject to random variation and will only represent an estimate of the true outcomes of the model. Repeating the simulation with the same input specifications will give different output results. The variability in the output depends on the number of histories that are simulated, and can be influenced favourably by using variance reduction techniques (Law and Kelton, 1982; Ripley, 1987; Kleijnen and Van Groenendaal, 1992).

Compared to the statistical-numerical approach, a single simulation run will take (much) more computer time, and the variability in the output implies that the likelihood methodology and standard optimization routines cannot be used. Testing the goodness of fit can be done by Chi-Square testing in which the variance of the simulated output is taken into account. Special stochastic optimization routines that allow for stochastic output are being applied for microsimulation disease models (Plaisier *et al.*, 1994; Seaholm, 1988), see Chapter VII for further discussion.

Despite these drawbacks, microsimulation has a number of advantages that make it worth consideration for analysis of cancer screening data. A model can be implemented in a computer programme within a relative short time, and modifications and extensions are in most cases also implemented easily. Apart from the computer time, there are no real restrictions with respect to the type of functional relations used in the model. For example, using a Weibull instead of an Exponential distribution requires only a minor modification. A distinct advantage is that the approach and the model that are used in data analysis can directly be used in predicting the effects and costs of screening policies. With the steadily increasing computing power the disadvantages of microsimulation will become of lesser importance since sufficiently large numbers of histories may be simulated within a reasonable time limits.

In the next section, I will give a brief description of our microsimulation programme MISCAN, which was developed on basis of our experiences with the SCRMOD macrosimulation model (Habbema *et al.*, 1978). Apart from MISCAN, one other microsimulation cancer screening model has been published (Parkin, 1985,1986).

MISCAN

The MISCAN simulation package has been developed both for analysis of data from screening projects, and for making predictions about effects (and costs) of alternative screening policies. The package consists of a coherent set of computer programs. The main program is the microsimulation program which can be used to simulate dynamic-stochastic models of screening for disease. See (Van Oortmarssen *et al.*, 1982) for details on the micro-simulation approach, and (Habbema *et al.*, 1984) for an overview of the possibilities of the programme.

The main simulation program offers a general framework for specifying the processes which describe the disease, the characteristics of the screening test(s), the impact of early detection on prognosis, and the various aspects of the screening programme. It can be used for simulating models of different levels of complexity, from very simple to quite complex. Because of the great variation in design and data-collection between screening projects, adaptations of the simulation program may be needed when a specific screening project is analyzed. These adaptations mainly concern the invitation system and the way in which data are reported. The program will then produce simulated output which conforms to the format of the results of the screening project.

The core of the MISCAN program; the natural history of the disease and early detection by screening has remained largely unchanged since the 1984 publication. One extension is used in Chapter III: the dwelling-times in a state can now be made age-dependent. For predictions about costs and effects of screening, the time of birth has been added to the life histories, which enables MISCAN to simulate a complete population and not only a single birth-cohort of persons. Some of the model-parameters, for example the cumulative incidence of the disease, can be made dependent on the birth cohort. Other parameters, especially those describing the screening programme, can change over calender time.

Further calculations, for example goodness of fit testing, or calculation of predicted costs, are made in application-specific post-processing routines which are presently often implemented in spreadsheet programs. In prospective evaluation the simulated screening results are combined with data on medical procedures (assessment, therapy), with data on costs, and when possible with utility values expressing the quality of life in a specific disease stage. The cost-effectiveness approach was first outlined in (Habbema *et al.*, 1987), and later applied to breast cancer screening (Van der Maas *et al.*, 1989; De Koning *et al.*, 1991; see section I.6) and cervical cancer screening (Koopmanschap *et al.*, 1990a; Van Ballegooijen *et al.*, 1990,1992,1993; see section I.7).

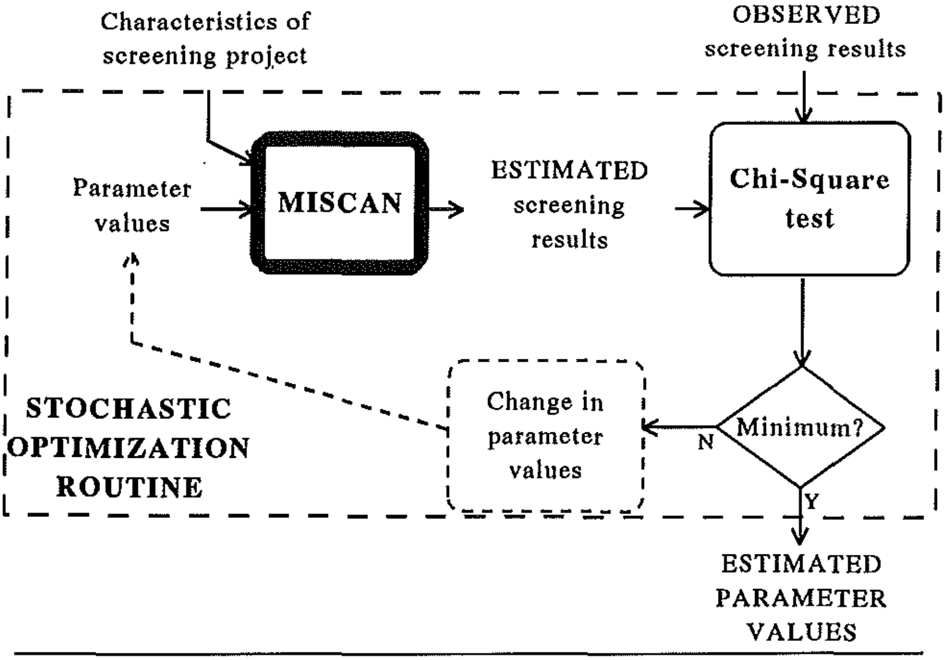
Characteristics of screening project
OBSERVED screening results
Parameter values
MISCAN
ESTIMATED screening results
Chi-Square test
Change in parameter values
N
Minimum?
Y
STOCHASTIC OPTIMIZATION ROUTINE
ESTIMATED PARAMETER VALUES

Figure Ie Analysis of screening data with the MISCAN model.

Parameter estimation

The estimation procedure for parameters of a MISCAN model follows the same steps that are used in statistical/numerical modelling (Figure Id). But adaptations are necessary to account for the stochasticity of the simulation (already discussed in section I.4), see Figure Ie.

For a given set of parameter values, MISCAN simulates screening results. The goodness of fit between these simulated results and the observed data is determined by Chi-Square tests that takes the variability in simulated results into account. Special optimization routines which allow for stochasticity of model outcomes may then be used to locate the combination of parameter values which minimizes the Chi-Square test. The dashed box "change in parameter values" indicates that we did not use fully automated iterative optimization routines in MISCAN applications thus far. In the analysis of the HIP study (Chapter II) a quadratic function is fitted to Chi-Square results of simulated parameter combinations, and new parameter values are chosen manually on basis of inspection of the fit results. This procedure can be regarded as a semi-automatic version of the Response Surface optimization method (Ripley, 1987).

Auxiliary programmes
In addition to MISCAN, we have developed and utilized a number of auxiliary models for analysis of data related to cancer screening. These models can be used to analyze a component of the model in detail, or can be regarded as simplification of a large part of the comprehensive MISCAN model.

*A*ge-*P*eriod-*C*ohort (APC) models have been used for detailed analysis of time-trends in mortality and incidence of a cancer. APC models are especially useful in case of cervical cancer, where there is substantial evidence for a cohort trend in incidence and mortality, probably related to changes in sexual behaviour and corresponding incidence of infections with certain strains of the Human Papilloma Virus (HPV) that have to be shown to be strongly associated with cervical cancer. Although APC models cannot be used to prove such a trend, they can serve to decompose the observed data into the three APC components on basis of presuppositions (Osmond and Gardner, 1982; Van Putten *et al.*, 1981; Moens and Van Oortmarssen *et al.*, 1987; Holford *et al.*, 1994). In case of cervical cancer, we used this model to derive cohort-specific relative risks on basis of mortality data (Habbema *et al.*, 1988; Van Ballegooijen *et al.*, 1993).

Other examples of detailed sub-models are the incidence-duration model which has been used for back-calculation of incidence rates for screen-detectable stages on the basis of (known) clinical incidence rates and (assumptions about) the duration of screen-detectable stages (Van Ballegooijen *et al.*, 1993), and a sub-model that has been used to test the consistency of age-specific data on clinical incidence, survival, and mortality.

The duration/sensitivity model developed by Walter and Day (1983) has been used to select starting values of parameters in the analysis of the HIP study data (see Chapter II).

Model-based analysis of breast cancer screening

MISCAN based-analyses have been carried out on data from the HIP study, from the two breast cancer screening projects in the Netherlands, and from the WE and Malmö trials in Sweden. The main objective of these analyses was to build and quantify a model which would allow us to translate the findings concerning reduction in breast cancer mortality of the randomized trials (HIP and Swedish projects) to other circumstances and to other screening policies, and in particular to the nationwide screening programme in The Netherlands. The breast cancer model conforms to the general structure presented in figure Ic, although the preclinical invasive stages are much more important than the non-invasive stages (e.g., ductal Carcinoma in Situ).

Before doing the model-based analysis, we analyzed the data from the HIP randomized trial study in detail. A copy of the statistical tape of the study was made available to us, containing datasets regarding participation and screening test results of the full study population and follow-up of all breast cancers diagnosed in the study group in which women were invited for screening and the control group of women who were not invited. We analyzed survival data from the screen-detected cases and the clinically diagnosed cases, investigating the potential impact of lead time bias and length bias (Habbema *et al.*, 1983).

Effectiveness of screening under age 50.

The issue whether the HIP study indicated that breast cancer screening is effective in reducing breast cancer mortality in women under age 50 is discussed in (Habbema *et al.*, 1986). Evaluation of the HIP data has given rise to much debate, especially about the effectiveness below age 50. This paper, and a paper from Prorok *et al.* (1981), were the first to challenge the prevailing opinion that the HIP data indicated that screening women under age 50 was hardly effective (Shapiro, 1977; Shapiro *et al.*, 1982). The main findings are given in Table I.3, Table I.3, showing a significant overall reduction in breast cancer mortality. Age-specific comparison suggests a beneficial effect of screening in all age-groups. However, the reduction is not significant in any of these groups. The beneficial effect, expressed in life-years gained by screening, is based on the assumption that the difference in number of breast cancer deaths stabilizes after the 14 years of follow-up represented in the data available at that time. Note that the highest gain is found in the youngest age group.

Table I.4 HIP study: effectiveness of breast cancer screening
Number of breast cancer deaths within 14 years after entry, in study- and control group, among cases diagnosed in first 7 years after entry; percent reduction in mortality with 95% confidence limits, and estimated number of life-years gained per 1000 women in the study group.

Age group (at entry)	Breast cancer deaths (numbers)		Mortality reduction		Estimated Life-years gained per 1,000 women
	Study group	Control group	%	range	
40-44	27	39	31	[-12,+59]	51
45-49	37	43	14	[-35,+46]	20
50-54	42	54	22	[-17,+50]	44
55-59	35	43	19	[-28,+49]	29
60-64	24	33	27	[-24,+59]	35
TOTAL	165	212	22	[+ 4,+37]	36

In a later article (Chu *et al.*, 1988), our finding of a homogeneous decrease in mortality across age-groups was reinforced. By using different and rather optimistic assumptions in statistical testing, Chu *et al.* demonstrated a significant ($p<0.05$) mortality difference for women who were below age 50 at entry.

MISCAN analysis of HIP study

The MISCAN analysis of the HIP study (see Chapter II) included both the DETECTION and the PROGNOSIS parts of figure Ic. A search was made for good-fitting values of the key parameters for both parts of the model. The HIP analysis resulted in estimates for the sensitivity of the two screening tests (mammography and physical examination), and also for the duration of the preclinical stages. An adequate fit of the HIP results was obtained, with two exceptions: the age-specific detection rates (see Figure If) and the incidence of interval cancers by time since negative last screening, see Figure Ig. In the latter figure, only one very strong discrepancy occurs: the model prediction of the incidence within the first 6 months since a negative examination is much too high. The discrepancy is at least partly due to the assumption of a constant sensitivity within a disease state. In reality, the sensitivity of clinical examination will probably be quite high in the months before a cancer is diagnosed on basis of clinical symptoms.

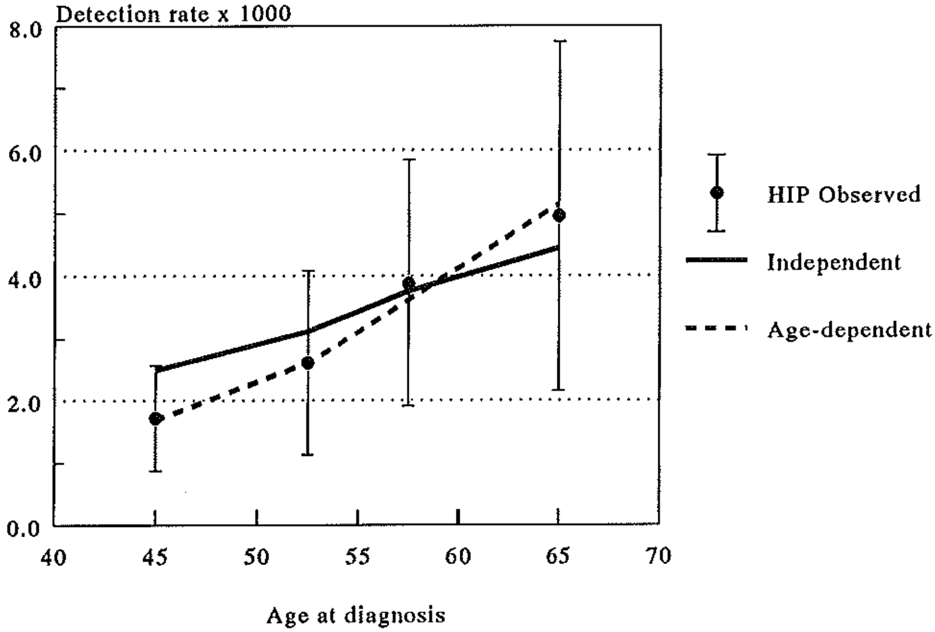

Figure Ig Detection rate at first screening. Comparison of the observed HIP results and predictions for model with and without age-dependent mean duration of preclinical stages.

MISCAN analysis of Dutch projects

The past decades showed tremendous improvements in mammographic techniques and a considerable shift in clinical stage distribution (see for example Coebergh, 1990). When we started the analysis of the costs and effects of nationwide breast cancer screening in The Netherlands, the HIP model had to be adapted in view of these trends. We analyzed the data from the screening projects in Utrecht (DOM project) and in Nijmegen, using the HIP model as a starting point, see Chapter III. In the HIP model, the clinical stage distribution was already explicitly included, and the quality of the mammography is represented by the sensitivity parameter. The possibility of finding tumours below 1 cm. in diameter was included by subdividing the stage "<20 mm." into two stages "<10 mm." and "10-19 mm.". Also, the stage intraductal carcinoma in situ was added, because of the increasing proportion of screen-detected cancers in this stage. The transition probabilities between preclinical stages were made age-dependent, which allows for age-dependent duration of the duration of the preclinical stage. Furthermore, the improvement in prognosis following early detection as specified in the HIP model had to be reconsidered because of these redefinitions of disease

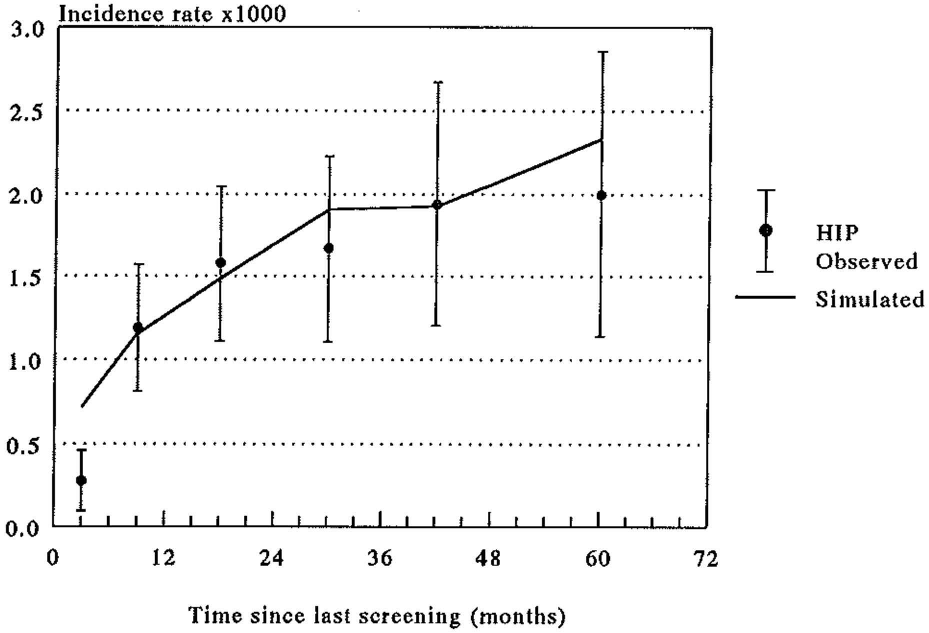

Figure If MISCAN analysis of HIP study: incidence of interval cancers by time since last negative screening test (per 1,000 women-years).

stages in the model.

We found a strong age-dependency of the duration of preclinical screen-detectable stages, the mean duration ranging from approximately 2 years at age 40 to 5 years at age 70, see Chapter III. Comparison of this model and the HIP model shows:

(1) If the same age-dependent duration of preclinical stages had been used in the HIP model, then the discrepancies between this model and observations from the HIP study would have been removed, see Figure If.

(2) The mean durations in comparable stages are indeed quite similar, reinforcing the validity of the model.

(3) The sensitivity of mammography shows a distinct improvement, see Table I.5.

Note that in the DOM project, which uses clinical examination in addition to mammography just like the HIP study, the incidence of interval cancers in the first half year is also lower than predicted by the model (see Figure IIIe in Chapter III).

Our findings about (age-dependent) duration have been used in explaining the accelerated rise in breast cancer incidence in the USA (Feuer and Wun, 1992).

Table I.5 Comparison of sensitivity of mammography in HIP model and in Dutch model.

HIP model		Dutch model	
diameter (mm.)	sensitivity	diameter (mm.)	sensitivity
< 20	28%	< 10	70%
		10-19	95%
≥ 20	56%	≥ 20	95%

Tumour doubling times based on repeated mammography in the Nijmegen project shows a age-dependency which is similar to our model assumptions: the median volume doubling times are 80 days for women younger than 50 years, and 188 days for women older than 70 years (Peer *et al.*, 1993). This paper reports a median volume doubling time of 157 days in the age group 50-70, which would indicate a median duration of 1.3 years for stage 10-19 mm. This is slightly longer than assumed in our model, which has an exponential distribution with *mean* duration of 1.3 years in this age-group, corresponding with a median duration of 1.1 years for the 85% of the cancers in this stage that have a transition to stage ≥20mm.

The analysis reported in Chapter III was confined to only part of the data from the Dutch projects. From the DOM project, only the first cohort invited in the city of Utrecht was included in the analysis, and from the Nijmegen project we only analyzed the first four screening rounds. In the mean time we arrived at a model that explains both the data from the first 6 screening rounds in Nijmegen and the additional data from the Utrecht region (De Koning *et al.*, 1990b,1991).

We did not use the findings concerning mortality reduction from the two Dutch projects, since these findings are based on case-control studies and are therefore possibly biased. Instead, we used the DETECTION model based on the Dutch projects to simulate the randomized trials in Sweden in order to analyze PROGNOSIS part of the model, i.e. the improvement in prognosis following early detection. Preliminary results, restricted to the Kopparberg/Östergötland trial, are mentioned in Chapter III.

If the differences between the Swedish projects in Kopparberg/Östergötland and Malmö regarding age-range, screening interval, and participation rate are taken into account, the model predicts almost equal mortality reduction percentages for the two projects. The difference in observed reduction is therefore interpreted as being random, and we derived an average effect by combining the results of these projects, weighting for the difference in size of the study populations (Chapter 7 in De Koning *et al.*, 1990b; De Koning *et al.*, 1991).

Prospective evaluation

In 1986, we were commissioned by the Dutch Health Insurance Council to perform an analysis of the effects and costs of nationwide breast cancer screening. The results were to be used in decision making about the implementation of the program, and in deciding about policy alternatives.

The model quantifications which resulted from the analysis of the screening projects have been used in predicting the results and effects of different policy alternatives and scenarios for the screening programme. Screening policies are simulated during a 27-year period, and compared with the situation of no breast cancer screening. Important output measures are the number of life-years gained, the increase in number of quality-adjusted life-years which summarizes the balance between favourable and adverse effects of screening, and the incremental costs per (quality-adjusted) life-year gained. In principle, these measures would enable comparisons to be made with similar figures for other investments in health care, provided that similar basic assumptions are being used.

The cost-effectiveness calculation have focused on the policy that is currently being implemented in the Netherlands, and in which women between ages 50 and 68 are invited for 10 screenings, i.e., a screening interval of 2 years. A comparison has been made with alternative policies in which women in the same age group will be invited with other intervals, or in which screening is extended to older and/or younger age-groups.

The reduction in total breast cancer mortality resulting from the main policy is predicted to be about 16% in the long term, or about 650 breast cancer deaths per year. The costs per life-year gained are 7,650 guilders (5% discount rate). Adjustment for changes in the quality of life does not lead to major changes. Compared to the negative effects, the positive effects other than gain in life-years are of the same magnitude. More details can be found in (Van der Maas *et al.*, 1989; De Koning *et al.*, 1991; De Haes *et al.*, 1991). In addition, calculations have been made regarding the impact of screening on demand for diagnostic and treatment facilities on national and on regional level (De Koning *et al.*, 1990a). Sensitivity analyses have been carried out for major uncertainties. The main uncertainty concerns the impact of early detection on mortality. On basis of the confidence interval derived from the Swedish trials, we predicted that the cost-effectiveness may vary between 4,500 and 22,000 guilders per life-year gained (De Koning *et al.*, 1991).

Similar MISCAN-based analyses have been carried out for other countries such as Australia (Carter *et al.*, 1993), Italy (Paci *et al.*, 1994), and Germany (Beemsterboer *et al.*, 1994).

Model-based analysis of cervical cancer screening

From a public health point of view, cervical cancer is less important than breast cancer in The Netherlands. Both the incidence and the mortality from breast cancer are approximately 10 times higher than from invasive cervical cancer. Evaluation of cervical cancer screening is more difficult, however, for a number of reasons:

(1) The natural history of cervical cancer is more complicated. Many of the pre-invasive lesions that are detected by the screening test (Pap-smear) will probably regress spontaneously, see Chapter IV. The long average duration of pre-invasive lesions means that although some lesions may develop rapidly, most of the progressive lesions that emerge in young women will only become invasive much later, see Chapter V. Thus, age is a key variable in analyzing cervical cancer screening data, whereas in breast cancer it is possible to analyze age-groups separately. A relatively minor complication is that women may have a hysterectomy for other reasons than cervical cancer, which protects them from developing invasive cancer. Modelling improvement in prognosis is relatively easy, since it can be assumed that treatment of screen-detected non-invasive cases will practically always result in cure.

(2) Significant incidence trends can be observed, related to both birth cohort en (calendar-) time.

(3) The effectiveness of cervical cancer screening has never been tested in a RCT. Still, there is quite convincing evidence for a strong effect on mortality. Comparison of Scandinavian countries with and without long-running nation-wide screening programmes shows a clear effect of screening (Läärä *et al.*, 1987, Day, 1984). Further evidence is supplied by case-control studies that were conducted to estimate the risk reduction accomplished by screening (Hakama *et al*, 1986; Van der Graaf, 1987).

(4) In many countries, including The Netherlands, organized and opportunistic screening co-exist. Usually, no complete registry exists of all smears and all detected cases, which seriously impedes data analysis and estimation of model parameters. And it is very difficult to make realistic predictions about alternative strategies for organized screening, as it is largely unknown how opportunistic screening patterns will adapt to different strategies.

Moreover, in recent years a strong association was found between presence of certain strains of the Human Papilloma Virus (HPV) and (pre-invasive and invasive) cervical cancer, suggesting a causal link between HPV infection and development of cervical cancer. HPV-based screening tests are available, and evaluation of cervical cancer screening now has to consider different alternatives for using cytologic and HPV-based tests, alone or in combination.

The effectiveness of cervical cancer screening can be analyzed by combining data regarding the "exposure" to screening with incidence and mortality figures. This may be done for a complete population, but it is more efficient to use a case-control design in which all *cases* (usually invasive cancers) are compared with a (stratified) sample from other women. The use of invasive cancers is not without problems however, since invasive cancers can be detected by screening. In the IARC/UICC study on the evaluation of cervical cancer screening programmes (Hakama *et al.*, 1986; IARC, 1987), in which we participated, a combined analysis was performed on data from screening programmes in 8 countries, see also section I.2. The goal of the study was to estimate the risk of cervical cancer associated with different screening intervals. The study compares the incidence of invasive cervical after one ore more negative Pap-smears with the incidence in unscreened women of the same age.

Table I.6 Relative protection against cervical cancer as a result of screening

The relative protection (RP) is given by time elapsed since last negative smear. Figures in parenthesis are No of women with cervical cancer. Estimates for British Columbia, and geometric mean for centrally organized screening programmes. Adapted from Tables II and III in (IARC, 1986).

Months since last negative smear	One previous negative smear		Two or more previous negative smears			
	British Columbia		British Columbia		All programs	
	RP	(N)	RP	(N)	RP	(N)
0-11	2.5	(10)	8.8	(7)	15.3	(25)
12-23	2.1	(10)	4.6	(7)	11.9	(23)
24-35	7.5	(2)	14.1	(1)	8.0	(25)
36-47	5.8	(2)	3.9	(2)	5.3	(30)
48-59	1.9	(12)	1.6	(3)	2.8	(30)
60-71			1.6	(5)	3.6	(16)
72-119					1.6	(6)
>119					0.8	(7)
Never screened	1.00		1.00		1.00	

Our contribution to this study was to make risk calculations for the British Columbia screening programme, see Van Oortmarssen and Habbema, 1986. The reduction in risk is expressed in terms of *relative protection*, i.e. the ratio of the incidence in unscreened women to the incidence in screened women. The findings for British Columbia and the average for all centrally organized programmes included in the study are presented in Table I.5. The data for British Columbia alone, although indicating a considerable reduction in risk for screened women, do not show a clear trend. This is caused by the low number of cases observed, even where the total number of women in the program is over 120,000. In contrast, the pooled results for the 5 centrally organized programs (including British Columbia) show a striking regular trend.

We used MISCAN for analyzing the data from the British Columbia cohort study (Habbema *et al.*, 1982c). In this analysis, many presuppositions were made, especially about the age-dependencies included in the model. We found that assuming a high proportion of regressive lesions would result in a good fit of the data, and we could not find a good fit without regression. However, because of the intricacy of this MISCAN cervical cancer screening model and the numerous pre-suppositions, it proved to be difficult to perform a formal test of hypotheses or to test the goodness of fit. We concluded that a better approach would be to start from a very simple model which could then be extended in successive steps. Although this could have been done with MISCAN, a statistical-numerical approach offers a number of advantages as discussed in section I.4. Therefore, I made another attempt to analyze the same screening data, with a model which can be regarded as a simplified version of the MISCAN model. The model, which is described in Chapters IV-VI, only covers the DETECTION box from the basic structure in Figure Ic. This implies that the mortality from cervical cancer and the impact of early detection on mortality are not included in the model. In Chapter VI a simple link between incidence and mortality is added to the model.

The model has 10 parameters (see Table IV.1, Chapter IV), which describe the incidence, the duration, and the progression of pre-invasive lesions, the relative risk of participants and non-participants, and the test-sensitivity of the Pap-smear. Mathematical formulas have been devised that express the clinical incidence and the detection rates at subsequent screening tests in terms of these parameters, see Chapter IV, appendix IVB. Numerical approximations of the formulas are used to calculate expected incidence and detection rates on basis of a given set of parameter values.

The model is fitted to the data from British Columbia using the procedure outlined in figure Id. Hypotheses are tested about regression of pre-invasive lesions and its possible age-dependency. It appears that a model with a high proportion of regressive lesions in women under age 34, and a high proportion of progressive lesions in older women, gives the best fit of the cohort study data.

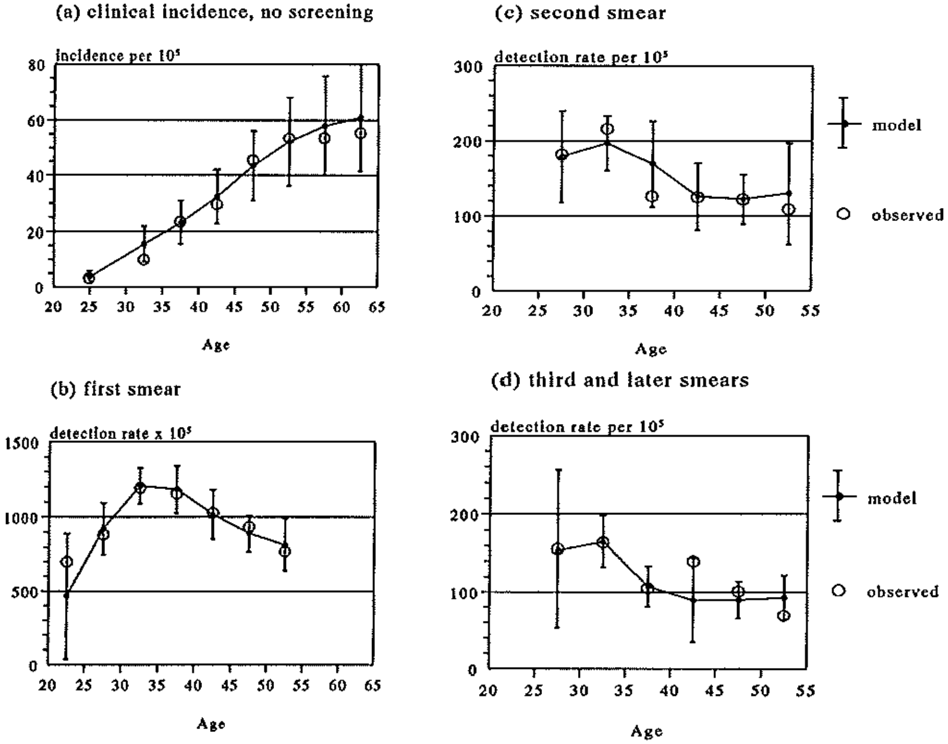

Figure Ih Comparison between observed results from British Columbia and model results concerning clinical incidence in unscreened women and detection rate at the first, second and third or later smears.

Figure Ih shows the striking agreement between model and data.

Estimates are derived for the duration of pre-invasive lesions, for the proportion regression among these lesions, and for the test-sensitivity of the Pap-smear. The combined confidence regions for these parameters as shown in fig VII.2 can be interpreted as follows: a higher proportion of progressive lesion will give a too high clinical incidence and the detection rates will become too high. The impact on clinical incidence is compensated in (a) and (b) by a lower total onset of pre-invasive lesions and a higher mean duration of pre-invasive lesions is higher, respectively. The impact on detection rates is compensated in (c) by assuming a lower sensitivity.

Table I.7 Summary of parameter estimates for cervical cancer screening models

Screening project	British Columbia	British Columbia	Dutch pilot regions	The Netherlands
type of model	simplified	MISCAN	MISCAN	MISCAN
	Van Oortmarssen (1991)	Habbema (1982c)	Habbema (1988)	Van Ballegooijen (1994)
mean duration of pre-invasive stages (years)	10-15	12-17	15-19	10-15
progression (%)	25-45	30-50	32-38	N.A.
sensitivity (%)	76-85	60-80	67-73	76-85

The estimated mean total duration of pre-clinical stages is 16 years. In the MISCAN analysis both the estimates for the mean duration and for the percentage of progressive lesions were slightly higher, see Table I.7. A larger difference occurs for the estimates of the sensitivity of the Pap-smear and also for the onset rate of pre-invasive lesions (1.7 per 1000 women-years under age 35).

A serious drawback of this simplified model is that the assumptions are in general very crude. For example, the age-dependency of the incidence of pre-invasive stages is modelled by distinguishing only two levels, for young women and for middle-aged women. At age 34, the incidence changes from $2.11\text{x}10^{-3}$ to $1.06\text{x}10^{-3}$, and the proportion of progressive lesions jumps from 16% to 60%. These sudden changes are not plausible, and one might wish to add additional assumptions to the model to make it more realistic. An alternative would be to assume other functional relations, e.g. a two-parameter Weibull distribution for describing the total onset as a hazard rate which decreases with increasing age, and a sigmoid type of function which describes the proportion of progressive cases as a function of age. But the fit of the present model is already very good, which means that additional data are required for discriminating between the present model and extensions. Another problem is that the complexity of the model equations will readily become too high, indicating that at some stage in the analysis, a switch from the statistical/numerical approach to a microsimulation model will become inevitable.

The statistical model has been validated against the outcomes of the IARC study concerning the relative protection following negative smears, see Chapter V. A mean duration of 16 years for the preclinical stage may seem incompatible with the more rapid decrease in relative protection following negative smears calculated in the IARC study (Table I.5). It appears that this apparent contradiction can be explained by acknowledging that women who had had a series of (negative) smears are likely to have more smears at which an invasive cancer can be detected. When available information about screening frequency is used in the model, then its predictions are in close agreement with the outcomes of the IARC study.

The consequences of these findings for the interpretation of the relative protection results published by the IARC study are discussed in Chapter VI. A considerable proportion of the screen-detected invasive cancers will be micro-invasive, and will on average have a much better prognosis than clinically diagnosed invasive cancers. Thus, the relative protection against incidence will be an underestimation of the reduction in risk of dying from cervical cancer following negative smears. In my opinion, the IARC recommendations of screening intervals of at most 3 years will lead to too frequent screening, since intervals of 5 years will still give 90% reduction in the risk of mortality from cervical cancer.

Further MISCAN-based analyses of cervical cancer screening have been carried out on the data from the screening program in three pilot regions in the Netherlands. A first analysis, reported in (Habbema *et al.*, 1985,1988) was performed in preparing the prospective evaluation of costs and effects of cervical cancer screening. The data from the pilot regions and the trends in the national cervical cancer mortality could be reproduced well. However, we encountered serious difficulties in fitting the data regarding cancers diagnosed outside the centrally organized screening program, presumably in part because of inconsistencies within these data. The main parameter estimates are quite similar to the values found in the analysis of British Columbia data, see Table I.7. A recent analysis, which concentrated on progressive lesions and the impact of screening, resulted in a good fit of the age-specific trends in the incidence of invasive cervical cancer and in cervical cancer mortality (Van Ballegooijen *et al.*, 1993)

Prospective evaluation

Although the same general (MISCAN-based) approach was used for prospective evaluation of screening for cervical cancer and for breast cancer (see section I.7), evaluation of cervical cancer screening proves to be more laborious (and challenging) because of the complications mentioned earlier in this section. The range of potentially (cost-)effective policies is very broad, and considerable differences in effectiveness exist between different screening ages or screening intervals. Many screening policies (differing in age-range, interval between screening, etc.) have been simulated to predict the beneficial and adverse effects, and the costs and savings of screening in the Netherlands during a 27 year period. For each policy, a comparison is made with the (highly hypothetical) situation in which no preventive Pap-smears are being made. The policies have been compared on basis of the effectiveness and the cost-effectiveness ratio. Systematic comparison of costs and effects of policies resulted in identification of a series of *efficient* policies, i.e., the policies that give maximum favourable effects at a given amount of incremental costs. The simulation results have been used to make recommendations about the ideal screening interval and age-range (Koopmanschap *et al.*, 1990a).

These efficient policies have been compared with the screening policy which had been used since 1976 in the Netherlands (7 invitations at 3 year intervals between age 35 and 53), with opportunistic screening, and with the combination of these two modes which is actually the situation which can be expected to occur. All three of them appear to be by far inferior to the efficient (organized) policies. Opportunistic screening is especially inefficient, and should in our opinion be discouraged. The cost-effectiveness of the existing organized policy with 7 invitations could be improved by approximately 50% by choosing a longer screening interval. A high attendance rate will have a favourable impact on the cost-effectiveness of screening (Koopmanschap *et al.*, 1990b). The first cost-effectiveness analysis pointed out that screening should not start at very young ages and suggested that screening should be extended to older age groups (55 years and over) (Van Ballegooijen *et al.*, 1990). This model had been fitted on available data in The Netherlands, with emphasis on the screening results in age-groups 35-54. The latter two conclusions, however, pertain to younger and older women, respectively.

It was decided to take a closer look at the data regarding younger and older women. After some adaptations, e.g. in the age-distribution of the onset of progressive pre-invasive lesions and in the age-dependency of survival of clinically invasive cancer, the Dutch incidence and mortality data of younger women could be fitted adequately. A special study has been carried out in which the incidence of invasive cancers in women between ages 55 and 68 has been related to the screening history of these women (Van der Graaf *et al.*, 1991). Model analysis of the outcomes of this study support our hypothesis of persisting onset of progressive pre-invasive lesions in these age-groups. Subsequent evaluation of the cost-effectiveness of screening policies reinforced our recommendations to extend screening to older ages, and to start between ages 25 and 30 and end around age 65 (Van Ballegooijen *et al.*, 1993).

These findings have influenced the recent policy modification in The Netherlands. Instead of 7 invitations between ages 35 and 53, invitations now start at age 30 and continue with 5-year intervals up to age 60. Another policy revision in the near future cannot be excluded, given the potential possibilities of the HPV-based screening test.

II. A MODEL-BASED ANALYSIS OF THE HIP PROJECT FOR BREAST CANCER SCREENING

Introduction

The HIP randomized trial was the first study to provide conclusive evidence of the effectiveness of annual mass screening for breast cancer (Shapiro *et al.* 1982; Habbema *et al.* 1986; Chu *et al.*, 1988). The HIP findings are supported by the results of recent non-randomized studies (Collette *et al.*, 1984; Verbeek *et al.*, 1984; UK Trial, 1988). Randomized trials are in progress in Sweden, the UK and Canada. For women in the age group 50-70, results from the Swedish 2-county trial confirm the HIP findings by showing a significant mortality reduction of 39% (Tabar *et al.*, 1985, 1989). The trials in Malmö and Edinburgh reported a smaller and non-significant reduction in mortality of about 20% in this age-group (Andersson *et al.*, 1988; Roberts *et al.*, 1990). None of these modern trials showed a clear mortality reduction for women below the age of 50.

In spite of the results of these trials, the choice of age-groups to be screened and intervals between screening examinations will still be difficult, since each trial only gives direct evidence concerning the screening policy used in the trial. Ideally, recommendations about breast cancer screening should be based on a cost-effectiveness analysis, in which the effects and costs of screening are estimated for different policies. These estimates should preferably be made with the aid of mathematical or simulation models. Important factors in these models are the *duration* of pre-clinical breast cancer, the *sensitivity* of the screening test(s), and the impact of early diagnosis and treatment on mortality (*efficacy*).

Models have indeed proven to be a valuable tool in the evaluation of screening for breast cancer (Eddy and Shwartz, 1982). Many of these models are primarily designed for the problem of finding effective screening policies (Knox, 1975), and not for a detailed analysis of results of ongoing screening projects. On the other hand, models can be found that are useful for analysis of (part of) the results of screening projects (Walter and Day, 1983). Like a few other models (Shwartz, 1978; Eddy, 1980), the MISCAN simulation program (Habbema *et al.*, 1984) can be used both for a detailed analysis of available data, and for subsequent prospective evaluation of screening policies. We now report the results of the MISCAN *analysis* of the data from the HIP study. In comparison with earlier model-based analyses of HIP data, this analysis is more detailed: the sensitivities of mammography and of physical examination are considered separately, a subdivision is made for tumour diameter, and the parameters describing duration, sensitivity and efficacy are analyzed simultaneously.

The central question addressed is:

which numerical assumptions for the *duration* of pre-clinical stages and *sensitivity* of the screening tests result in good agreement between model-based simulated results and observed results of the HIP study?

Material

The HIP study for the early detection of breast cancer started in 1964. About 60,000 women aged 40-65 were randomly assigned to the study group or to the control group (Shapiro and Venet, 1966). Women in the study group were invited for an initial screening. Participants in the first round were offered 3 additional yearly screening examinations (in practice, the average interval between screenings was about 13 months). Two screening tests were applied at each examination: mammography and physical examination. All new breast cancer cases in study and control groups have been recorded during a period of 10 years since entry in the study, and are followed-up for mortality. Results for 14 - 18 years of follow-up have been published (Shapiro *et al.*, 1982, 1988; HIP Study, 1981 ; Chu *et al.*, 1988).

The present analysis uses data from breast cancer cases detected within 7 years after entry into the study, with a follow-up of 14 years. Thus, all breast cancer cases diagnosed within about 3½ years after the final screening round in the study group are included. This "catch-up" period is needed to ensure that the control group counterparts of the screen-detected cases in the study group will (nearly) all be diagnosed. Without such a catch-up period, the effect of screening on mortality would be underestimated (Habbema *et al.*, 1986).

Model assumptions are tested by comparing observed results from the HIP study with results obtained by simulation using specific assumptions about the model-parameters. An overview of the HIP data used is given in Table II.1, detailed information is given in Appendices IIA and IIB.

The 443 cases of breast cancer diagnosed in the control group within the 7 year study period have been classified by age at diagnosis. The 425 study group cases have been grouped according to the mode of detection: diagnosis as a result of screening (132 cases), interval cancers diagnosed after a negative screening result (173 cases), or cancers diagnosed in women who did not participate in the first screening round (120 cases). The screen-detected cancers and the interval cancers are classified by screening round, age at diagnosis, and tumour diameter (under or over 2 cm). Screen-detected cases are also classified by the mode of detection (mammography alone, physical examination alone, or both tests positive). Interval cancers are also classified by the time since the last negative test result.

Table II.1 Overview of observed data from the HIP study that have been used in testing of model-assumptions.
All data were derived from the Statistical Tape of the HIP study. Results of the testing procedure are indicated for the standard model: No mark: $p>.05$, *: $0.01<p<.05$, **: $p<.01$

A. HIP Control group results.
1. Incidence by age
2. Case-mortality by age

B. HIP Study group: screen-detected breast cancers.
1. Detection rate by age at diagnosis, first round.
2. Detection rate by age at diagnosis, rounds 2,3,4.
3. Detection rate by screening round (2,3,4)
4. Distribution according to modality, by screening round (1,>1)
5. Distribution according to tumour size, by screening round (1,>1)

C. HIP Study group: Interval cancers.
*1. Incidence by age at diagnosis.
**2. Incidence by time since last screening
3. Incidence by round of last screening
4. Distribution according to tumour size.

D. HIP: Difference in breast cancer mortality between study group and control group.
1. Mortality difference in absolute numbers, by time since entry.
2. Difference in case-mortality, by time since entry.

The breast cancer mortality in study and control groups in the HIP study is analyzed by age at entry and by years since entry. Both absolute numbers of deaths and case-mortality rates are considered. Standard actuarial techniques have been used in calculating the case-mortality, thus correcting for death from other causes and for cases that are lost to follow-up.

Survival figures, based on 7 to 14 years of follow-up, are used directly to model the survival of control group cases. Breast cancer incidence rates for the US (Third National Cancer Survey, 1975) are used to extend the model to age-groups that were not included in the HIP study. Within the age-range of the HIP study, the incidence in the HIP control group appears to be some 13% higher than the TNCS incidence (US totals). Breast cancer mortality rates in the HIP control group are also a little higher than rates reported from New York (NCHS, 1980).

Method

The MISCAN computer program (Habbema *et al.*, 1984) uses micro-simulation, *i.e.*, simulation of individual life-histories of (fictitious) persons. The same persons are considered twice. First, in the absence of screening, they add up to an unscreened population, *e.g.*, the HIP control group. Next, screening is simulated. Some of the life-histories will change as a result of screening. The life-histories then add up to a screened population, which for example may be compared with the study group of the HIP study.

The life-histories are simulated according to the specifications of a breast cancer epidemiology and screening model. Some of the model parameters, like the age-distribution of the screened population, and the survival of clinical stages of breast cancer, can be assessed directly from available data. Simulations are carried out to test assumptions for parameters for which direct estimation is not possible (Figure IIa). These simulations will produce simulated results of the HIP study. Goodness-of-fit statistics (Chi-square, Student t) are used for comparing the simulated results with the observed HIP data.

The model

Population and screening policy

The age distribution of the simulated population at first screening is obtained from the HIP cohort. The life table used in the model (death from causes other than breast cancer) is based on the New York State life table 1969-1971 (NCHS, 1980).

Attendance at first screening is taken to be 66.8% for all age-groups. This corresponds to the level in the HIP study, in which no clear age-trend was observed. The participants in this first round were invited for 3 more screenings. In modelling their attendance pattern, the observed strong association between participation rates in subsequent rounds has been taken into account. The interval between screenings is assumed to be 1.08 years for all women and all rounds. The resulting simulated participation rates in rounds 2-4 differ less than 1% from the observed rates.

The observed breast cancer incidence rates in the control group and in non-participants have been employed in estimating the relative risks of participants (RR=1.1) and non-participants (RR=0.8).

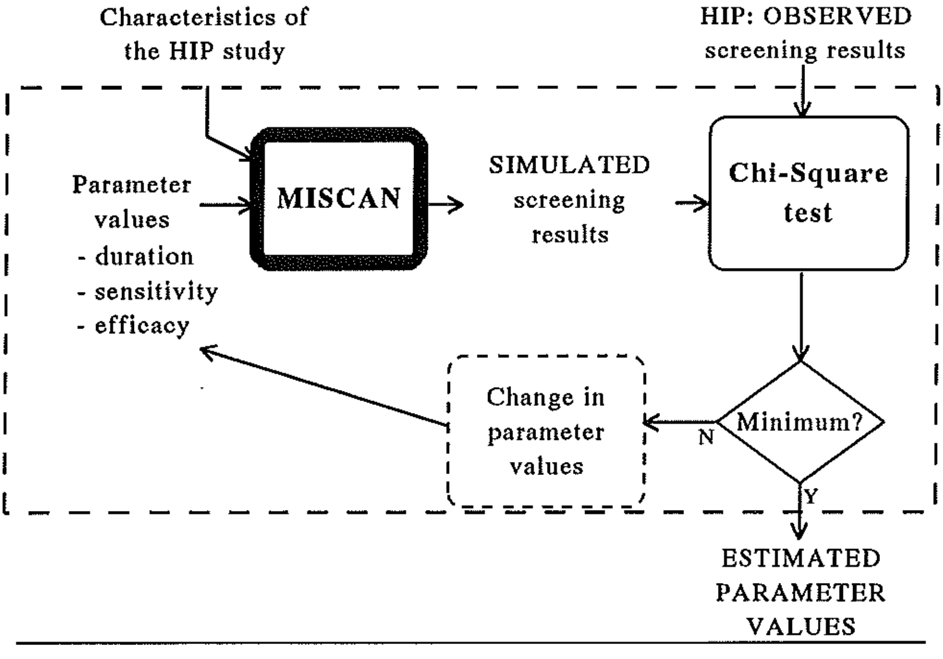

Figure IIa Use of the MISCAN simulation program in finding model assumptions which result in a good fit of the results of the HIP study, by testing simulated results against observed results.

Disease model

The natural history of the disease is modelled as a progression through a number of states, starting from state NO BREAST CANCER (Figure IIb). The upper part of Figure IIb describes the course of the disease in the absence of screening. The boundary between NO BREAST CANCER and pre-clinical disease is defined in terms of detectability at a screening examination: the screen-detectable stage is entered when the tumour becomes detectable by the screening modalities used. For the HIP study, the critical tumour diameter is about 1 cm.

The pre-clinical screen-detectable phase is subdivided in 2 states, according to tumour diameter. Variability in the growth rate between tumours is represented by an exponential probability distribution of the dwelling times in the pre-clinical states. A high correlation between the durations in the 2 states is used to reproduce approximately constant growth rates of individual tumours. Thus, the duration of pre-clinical breast cancer is fully described by the 2 mean dwelling times in the pre-clinical states.

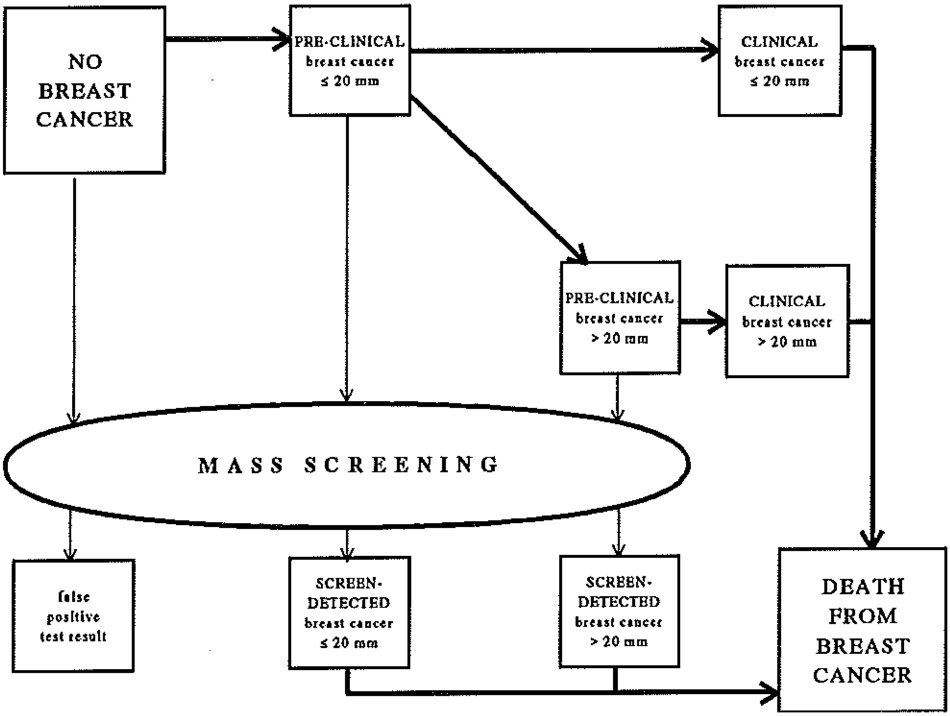
NO BREAST CANCER
PRE-CLINICAL breast cancer ≤ 20 mm
CLINICAL breast cancer ≤ 20 mm
PRE-CLINICAL breast cancer > 20 mm
CLINICAL breast cancer > 20 mm
MASS SCREENING
false positive test result
SCREEN-DETECTED breast cancer ≤ 20 mm
SCREEN-DETECTED breast cancer > 20 mm
DEATH FROM BREAST CANCER

Figure IIb Structure of the disease model for breast cancer.

From the HIP control group, it is estimated that in the absence of screening, about 30% of breast cancers are diagnosed before the tumour is 2 cm in diameter. This corresponds with transition to the state CLINICALLY DIAGNOSED BREAST CANCER ≤ 2 CM. The other 70% first proceed to state PRE-CLINICAL BREAST CANCER > 2 CM, and then to CLINICALLY DIAGNOSED BREAST

CANCER > 2 CM. After diagnosis and treatment, women are at risk of dying from breast cancer and also run the same age-dependent risk of dying from other causes as women in NO BREAST CANCER and in pre-clinical states.

In the case of early detection women will enter one of the two "SCREEN-DETECTED" states in Figure IIb.

Efficacy

The efficacy of early detection is expressed as a change in the age of death from breast cancer in screen-detected cases. There are 3 possibilities: no difference in time of death, delay, and cure. A 4th possibility, earlier death as a result of early detection and treatment, is assumed to be negligible. The probabilities of cure and delay, and the distribution of the delay time, are estimated from the Relative Case Fatality Rates (RCFR's). The case fatality is the complement of case survival, and has been calculated by standard actuarial methods, again correcting for patients who are lost to follow-up and patients who die from other causes. The observed RCFR is calculated from the 132 screen-detected cases in the HIP study group. The expected RCFR of these cases in a situation without screening is computed from the survival data of control group cases, with a correction of 2 years for the lead-time bias (Habbema *et al.*, 1983)

The ratio of observed to expected RCFR is used as a measure of the efficacy of early detection and treatment, see Figure IIc. A quadratic curve was fitted to the ratios in years 3-14, and extrapolated to a full 20 year follow-up period. Just after diagnosis, the ratio is about 30%, representing the breast cancer cases that do not benefit at all from screening. In the long run, the ratio approaches 70%. The other 30% represents the proportion of cure, *i.e.*, women who do not die from breast cancer after early detection by screening, but who would have died from breast cancer in the absence of screening. The remaining 40% represents women in whom breast cancer death is delayed as a result of screening.

These efficacy estimates apply directly to a model with a single pre-clinical state. For the 2-state model, the probability of cure will be highest for state $\leq$2 cm. With cure rates of 40% and 10% for earlier and the later states, respectively, the overall cure rate of 30% is maintained. It is assumed that both pre-clinical states have the same probability of delay (40%).

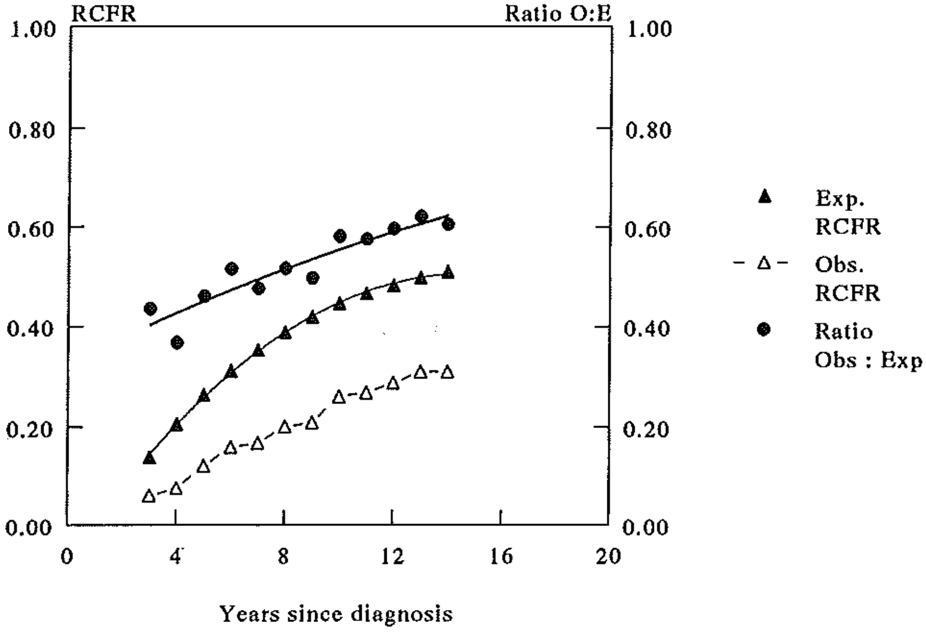

Figure IIc. Observed relative case fatality rate (RCFR) for the 132 screen-detected breast cancers in the HIP study, and expected RCFR in the absence of screening, obtained from control group cases by correcting for lead time. A quadratic curve has been fitted to the ratio: Observed RCFR / Expected RCFR .

Basic parameters

There are statistical and practical limitations to the number of parameters that can be varied in the analysis of the HIP data. Therefore, the analysis was confined to 4 basic parameters: the mean *duration* of the 2 pre-clinical states, and the *sensitivity* of the screening tests in these states. These 4 parameters have been assumed not to depend on age.

Results

Duration and sensitivity

Combinations of assumptions about the mean duration and the sensitivity of the screening tests in the 2 states have been tested. Values for the 4 basic parameters can be arranged in a 4-dimensional grid. A search procedure was used to identify the area of HIP-compatible assumptions in this grid, *i.e.*, assumptions for which the simulated results show a good fit with the observed results of the HIP study.

The best-fitting values are presented in Table II.2. Within the 2 states, quite large variations in assumptions are allowed: the mean durations in each state may vary between 1.0 and 1.5 years. The best-fitting value for the sensitivity in state ≤ 2 cm is 55%. Although this is lower than the value of 85% found in state >2 cm, much variation around these values is possible and the areas of good-fitting values show a considerable overlap. The allowed variation in the total mean duration and the average sensitivity is smaller. Figure IId shows the area of HIP-compatible assumptions. The average sensitivity can vary between 50% and 80%, and the total mean duration between 1.6 years for high sensitivity values and 2.7 years for low sensitivity.

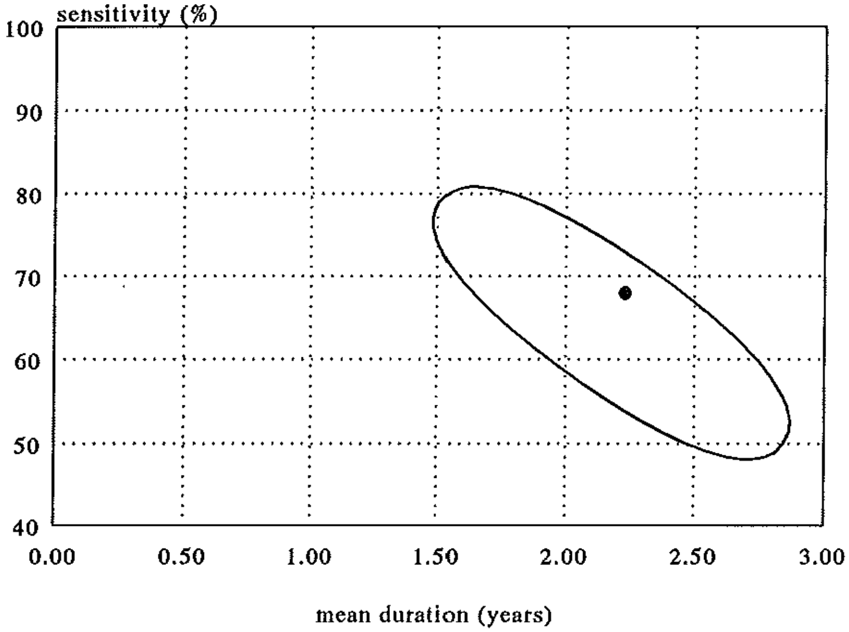

Figure IId. MISCAN Analysis of HIP Study. Area of combinations of assumptions about average sensitivity and mean total duration in the pre-clinical states that are in agreement with results of the HIP study.

Table II.2 Values of basic parameters of the standard model.

PARAMETER	value
Mean duration	
pre-clinical tumour, diameter ≤ 2 cm	1.25 years
pre-clinical tumour, diameter > 2 cm	1.4 years
mean total pre-clinical duration	2.23 years*
Sensitivity of combined screening tests	
pre-clinical tumour, diameter ≤ 2 cm	55% (MGM 28%, PE 38%)
pre-clinical tumour, diameter > 2 cm	85% (MGM 56%, PE 66%)
mean sensitivity	68%**

* only 70% of tumours will pass through both states. ** average based on stage distribution in first round. MGM - mammography, PE - physical examination

The best fitting combinations of assumptions are found in a narrow range around what we will call the "standard model", which has an average sensitivity of 68% and a total mean duration of 2.2 years (Table II.2).

Comparison of simulated and observed results

The fit of the standard model is good for all data ($p>0.05$), except for interval cancers by age ($0.01<p<0.05$) and interval cancers by time since last negative screening ($p<.01$) (Table II.1).

The simulated age-specific breast cancer rates for the control group are based directly on the observed rates. Therefore, finding a perfect fit is hardly surprising. Figures IIe and IIf show the results for the screendetected cases in the first round and for the interval cancers. The simulated rates in the 2 figures are similar, and reflect the incidence in the control group. This is a direct consequence of the assumption that neither the duration of pre-clinical breast cancer nor the test-sensitivity depend on age.

The fit for the trend in detection rates per round is good. A statistically significant discrepancy between simulated and observed results for interval cancers by time since last negative screening occurs in cases diagnosed within 1/2 year: the simulated rate is too high. This discrepancy may be due to the assumption that the sensitivity is constant within each state. In reality, the sensitivity can be expected to be higher shortly before symptomatic diagnosis of breast cancer, and lower just after entering the screen-detectable state.

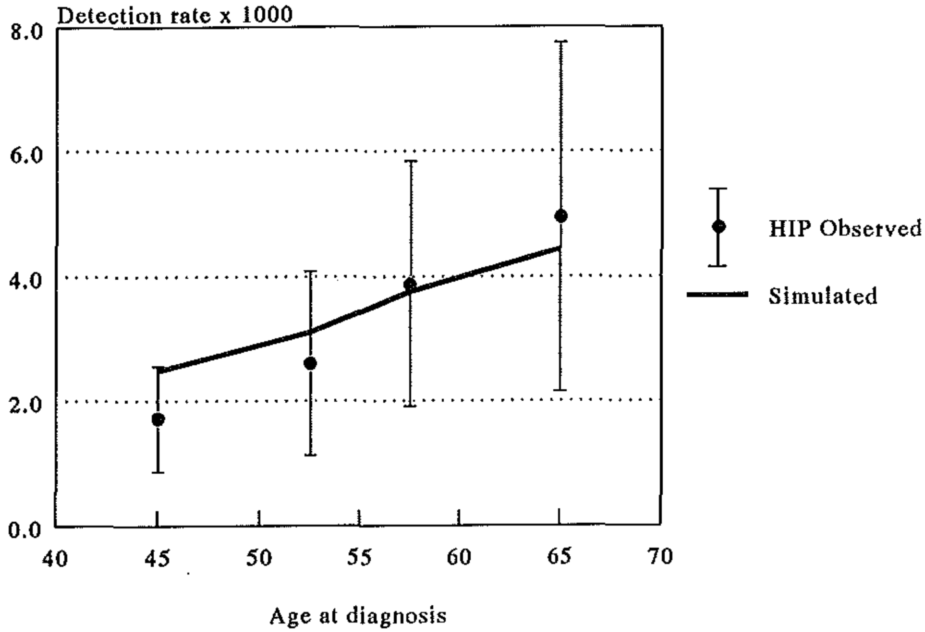

Figure IIe. MISCAN Analysis of HIP Study: Comparison of observed and simulated results of the standard model. Screendetected breast cancers per 1000 women examined in the first round, by age at diagnosis. Dots and vertical lines represent observed results with 95% confidence intervals; the continuous line connects the simulated rates.

The tumour size distribution is most favourable for the cancers detected in the repeat rounds. The observed proportion of tumours smaller than 2 cm is just over 40% for the interval cancers, while for the cancers detected in the first round, the simulated proportion is more favourable (53%) for screen-detected (first round) cases (Table II.3). Both the simulated and the observed proportion of interval cancers below 2 cm are higher than the observed proportion (30%) in the control group.

Table II.3 MISCAN Analysis of HIP Study: Comparison of observed and simulated results of the standard model. Percentage of tumours with diameter ≤ 2 cm, by mode of detection.

MODE OF DETECTION	Simulated: ≤ 2 cm	Observed: ≤ 2 cm
Screen-detected, round 1	53%	41%
Screen-detected, round 2-4	69%	67%
Interval Cancers	38%	43%
Control Group	30%	30%

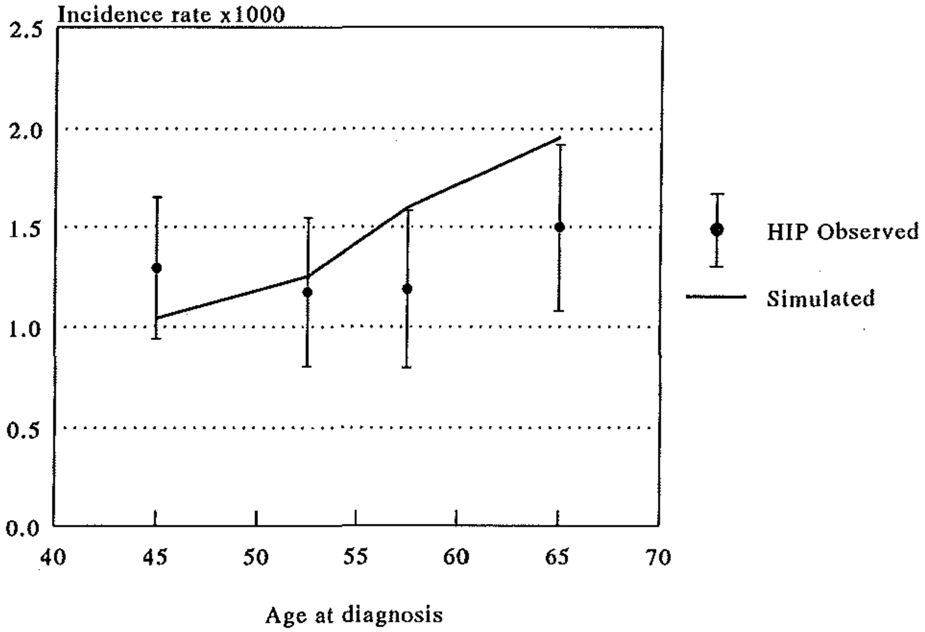

Figure IIf. MISCAN Analysis of HIP Study: Comparison of observed and simulated results of the standard model. Breast cancers detected after a negative screening test, per 1000 woman-years since last examination, by age at diagnosis. Dots and vertical lines represent observed results with 95% confidence intervals; the continuous line connects the simulated rates.

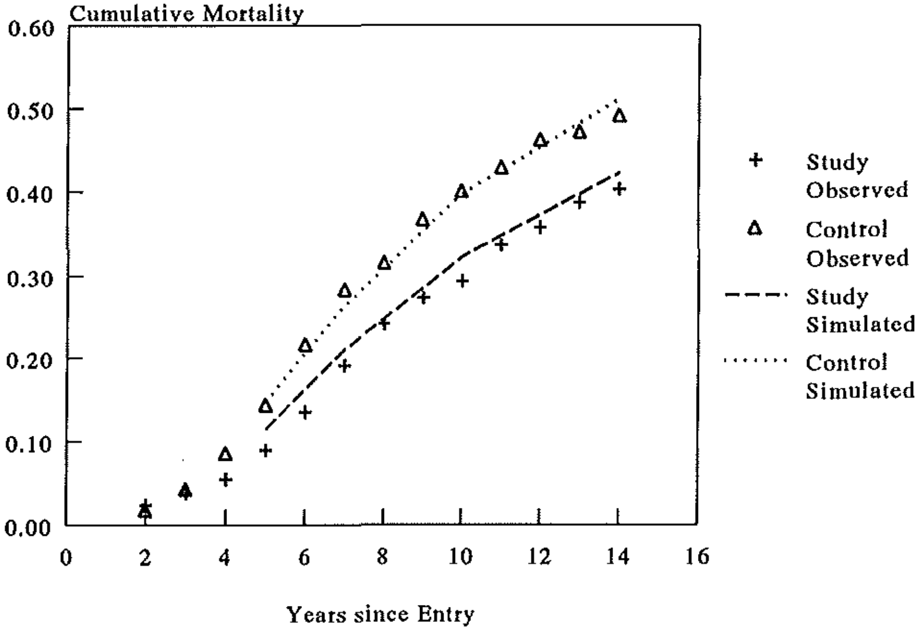

Figure IIg. MISCAN Analysis of HIP Study: Comparison of observed and simulated results of the standard model. Case-mortality in study and control group, by time since entry into the study. Lines connect simulated rates, points represent observed rates.

Case-mortality is the ratio between number of deaths and number of diagnosed breast cancers in which the time of entry into the study is the starting point (Breslow *et al.*, 1981). In Figure Iig, the simulated difference in case-mortality between the two groups is compared with the observed difference. In the first years, the simulated difference is smaller than observed. This discrepancy cannot be removed within the present framework of the model. The fit for 8 to 14 years of follow-up is adequate, but not perfect. It could have been improved by calibration of model parameters concerning efficacy of early detection. At 14 years since entry, both the simulated and observed differences in case-mortality rate or equal to 0.09. The same difference of 0.09 was predicted for 18 and 20 years of follow-up. Published results for 18 years of follow-up (Shapiro *et al.*, 1988) show a slightly higher difference of 0.107 in case mortality rate.

Discussion

The model gives an adequate fit of the HIP results for a mean pre-clinical duration of between 1.6 years and 2.7 years, and a sensitivity for the screening test of between 50% and 80%.

Simplified model

In the exploratory stages of the analysis, simplified models have been used, for example a model with a single pre-clinical state in which the screening examination is not separated into 2 modes of detection (mammography and physical examination): the *1-state/1-mode model.* This model differs from the 2-state/2-mode model discussed in the previous sections in 2 (related) respects: it has fewer parameters and, most importantly, less HIP results can be used in testing the parameters. Different types of probability distributions for the duration of the pre-clinical screen-detectable state were tested. Good results have been obtained with the exponential distribution. These could not be improved upon by using more complex distributions. The same conclusion has been drawn using a similar but simplified model (Day and Walter, 1984). Therefore, the exponential distribution has been used subsequently in the 2-state/2-mode model.

The area of HIP-compatible assumptions for duration and sensitivity is much larger for the 1-state/1-mode model than for the 2-state/2-mode model. In the 1-state/1-mode model, values of the sensitivity between 80% and 100% are not rejected, see also Walter and Day (1983) and Day and Walter (1984), but these values clearly appear to be incompatible with HIP results in the more detailed 2-state/2-mode model.

Sensitivity

In the present analysis, it is assumed that false negative test results are completely random, *i.e.*, that there is no association between results of mammography and physical examination, and no association between test results at different examinations. This is probably unrealistic, since some tumours may be more difficult to detect than others. Introduction of (some degree of) systematic false test results would result in somewhat lower values for the combined sensitivity than we have found.

Goldberg and Wittes (1978) assume, like us, that the results of mammography and physical examination are independent. They consider the modality distribution of screen detected cancers in the HIP study, and estimate a range of 0.44-0.76 for the sensitivity of the combination of the 2 tests, which is comparable to our range of 0.50-0.80. Sensitivity values below 0.50 would give rise to too high rates of interval cancers, these are not considered by Goldberg and Wittes.

Duration of pre-clinical stages

Previous duration estimates are based on a sensitivity of 100% (Shapiro *et al.*, 1974; Zelen and Feinleib, 1969), or use assumptions that lead to high sensitivity values (Eddy, 1980). Because of the inverse relationship between sensitivity and duration values, see Figure Iid, it is clear that this leads to too short an estimated duration of pre-clinical stages. The high value (2.4 years) found by Zelen and Feinleib (1969) fits within our range of 1.6-2.7 years, but it is based on early, incomplete results of the HIP study; use of their method on complete results would result in a much lower value.

The mean duration of about 1.0-1.5 years for the state 1-2 cm in our 2-state/2-mode model is compatible with clinical studies which show great variation in growth rates, but on average suggest a tumour diameter doubling time of about 1 year.

The simulated rates in Figure Iie for screen-detected cancers in the first round are too high at younger ages, and too low at older ages. The same deviations appear in screen-detection rates for repeat rounds (not shown). Complementary differences are found for interval cancers (Figure Iif). This could be expected since the total number of cancers is related to the good-fitting control group incidence. In the present model, both sensitivity and duration of pre-clinical breast cancer are independent of age. A longer duration and/or a higher sensitivity with increasing age could diminish the discrepancies in the youngest and the oldest age-groups.

Other models

The model of Shwartz (1980,1981) is based on the concept of tumour growth rates, using the average doubling time of 1.1 years reported by Kusama *et al.* (1972). It leads to good agreement with HIP data.

Very different results are obtained in the model of Eddy (1980), who uses the so-called "progression assumption", implying that once a cancer becomes detectable by a screening test, the sensitivity is 100%. The assumed mean duration of screen-detectable disease of about 1.4 years results in a good explanation of HIP data. Note that this duration compares well with the product of sensitivity and duration (0.68 × 2.23 = 1.52, Table II.2) in our model.

Efficacy

The discrepancy between simulated and observed HIP case-mortality in the first years of follow-up is noteworthy. The more slowly evolving difference between study and control group in the model simulation results is in agreement with the findings from recent trials in Sweden and in the UK (Tabar *et al.*, 1989; Andersson *et al.*, 1988; Roberts *et al.*, 1990). The early reduction in mortality in the HIP study could possibly be ascribed to chance variation.

The overall reduction in mortality in the HIP study group is statistically significant (Shapiro, 1977), but the confidence range (1%-34% reduction) reflects the still great uncertainty about the amount of reduction. No clear difference in reduction can be observed between age-groups (Habbema *et al.*, 1986). Therefore, in the model, the efficacy of early detection is assumed to be independent of age. This assumption is supported by the results of a statistical reanalysis of the HIP study of Chu *et al.*, 1988; who demonstrate a significant mortality reduction in women below age 50 at entry in addition to the already established significance of the mortality reduction in the higher age-groups. However, these favourable results in women under 50 have not (yet) been reproduced in the recent trials.

Extensions to the model

Further extensions of the model have not been explored, because of the limited number of cases in the HIP study. The number of breast cancers is already small for some of the classifications used, especially for screen-detected cases and for deaths from breast cancer. Missing data inhibit the use of other subclassifications, *e.g.* by lymph node status.

A related issue is the validity of some of the assumptions used in the model. The assumptions, such as the independence of test results, the correlation of the duration in the 2 pre-clinical states, or the age-independence of the mean duration of pre-clinical states, could have been varied easily in the MISCAN computer programme. Again, the limited number of cases in the HIP data was the main reason that this was not done.

Other extensions of the model, *e.g.*, radiation risks and false positive test results, have been left out because they are more easily analyzed separately, and do not interfere with the estimation of the basic parameters of the model. For example, the specificity of the screening tests can be directly estimated from follow-up data of women with positive test results.

Modern mammography

An adapted version of the model has been used to analyze the results of more recent screening projects (Collette *et al.*, 1984, Verbeek *et al.*, 1984) in the Netherlands. Mammographic techniques have improved since the HIP study. The lower limit of screen-detectable tumour-diameter has decreased from about 1 cm to about 0.5 cm. Cases of intraductal carcinoma in situ are also increasingly detected at screening. Therefore, the model has been extended to a 4-state model (dCIS, and tumour sizes ≤1 cm / 1-2 cm / >2 cm). The analysis indicates that the results concerning the average duration of the states are in agreement with the results of the analysis of the HIP data, at much higher values for the sensitivity of mammography (Chapter III). Furthermore, it was necessary to include age-dependence of the mean duration in the model, with a mean duration at age 65 being about twice the mean duration at age 40. When transferred to the HIP model, this same age-dependency would practically remove the discrepancies found in Figures IIe and IIf, and Figure If in Chapter I. These findings show the possibilities of models for comparing screening programmes, by explaining differences and similarities in results between these programmes.

Application of the model

Models for screening of breast cancer can be used to predict effectiveness of screening policies, and in making cost-effectiveness calculations regarding the age-range to be screened, the interval between screening examinations, and the type of screening test(s) to be used. Test-sensitivity, duration of pre-clinical stage, and efficacy of early detection are crucial parameters in comparing different policies, especially in relation to the screening interval.

The extended (4-state) model has been used in a comprehensive evaluation of public health effects and costs of nation-wide breast cancer screening in the Netherlands. The impact of screening on the demand for health care facilities was also predicted (Van der Maas *et al.*, 1989; De Koning *et al.*, 1990). This approach to prospective evaluation of screening policies can be equally useful for other countries.

Appendix IIA.

HIP BREAST CANCER SCREENING MODEL: SIMPLIFYING ASSUMPTIONS AND DIRECT ESTIMATES

Age-dependency

The natural history of the disease is assumed not to depend on age. The onset rates (the transition from NO BREAST CANCER to the first preclinical screen-detectable state), do increase with age.

Onset

Apart from a difference in risk of breast cancer between refusers and participants, no other risk categories are considered in the model. As a consequence, women who attend all four screening rounds have the same risk as women that only attend the first screening. Shapiro *et al.* (1982) estimated relative risks of participants and non-participants from the observed clinical incidence in the control group and in the refusers (non- participants to the initial screening) in the study group. In the period until seven years after entry, the observed incidence rates are 2.12 and 1.75 per 1000 woman-years, respectively. The estimated expected incidence for participants is then 2.31 per 1000 woman-years, and the risk ratio of participants to non-participants is 2.31 : 1.75. This ratio is used in specifying the preclinical incidence, i.e. the transition rate to the preclinical screen-detectable state, for participants and non-participants in the simulated cohort.

In testing model assumptions against HIP results, different assumptions are made on the total duration of the preclinical states. Depending on the mean total duration, the age-specific preclinical incidence is approximated from clinical incidence data on breast cancer, by putting forward this incidence by the assumed mean duration. For example, if the mean duration of the preclinical states is assumed to be 5 years, then the preclinical incidence can be approximated from clinical incidence figures in 5-year age-groups by moving up all age-groups by one category.

Preclinical disease

Duration in the two preclinical states is assumed to be 100% correlated. The mean durations in each state are essential parameters of the model.

The observed clinical tumour size distribution of the HIP control group (30%/70% for cases with known tumour size) is used directly in the model: from the state PRECLINICAL ≤ 2 CM., 30% will proceed to CLINICAL ≤ 2 CM., and 70% will proceed to state PRECLINICAL >2 CM. and will be diagnosed in state CLINICAL >2 CM.

Survival

The transitions from states CLINICAL to DEATH FROM BREAST CANCER are parametrized directly from the observed distribution of the survival time after clinical diagnosis in the HIP control group (relative survival, up to 14 years of follow-up).

Sensitivity

The results of mammography and of clinical examination for preclinical breast cancer cases are considered to be independent. Furthermore, in varying the sensitivity values it is assumed that a constant difference exists between the sensitivity of mammography alone and the sensitivity of clinical examination. From the distribution of screen-detected cases by mode of detection (33% mammography alone, 45% clinically alone, 22% both modalities positive, it is estimated that the percentage of cases missed is 10% higher for mammography than for clinical examination. Although the data might suggest an age-relationship for this difference, this is neglected for the time being because it is based on rather small numbers: from a total of 31 cases detected below age 50, 19 were detected by clinical examination alone, and only 6 by mammography alone. After fixing the sensitivity difference at 10% for both preclinical states, only one unknown parameter remains that describes the quality of the screening test in a preclinical disease state: the sensitivity of the combined screening test.

Attendance

Table IIA.1 Participation in the HIP study, and simulated rates that are based on the model specifications concerning attendance.

ROUND	Participation	
	OBSERVED	*SIMULATED*
1	67%	*67%*
2	53%	*53%*
3	49%	*49%*
4	46%	*45%*

Appendix IIB.

OBSERVED HIP RESULTS AND SIMULATED OUTCOMES.

Table IIB.1 Clinical incidence and mortality of breast cancer, adapted to level in HIP control group from TNCS (1975) and the New York State Life Table, NCHS (1975).

Age group	Clinical Incidence		Mortality	
	TNCS	*simulated*	New York	*simulated*
25-29	0.10	*0.13*	0.02	*0.03*
30-34	0.26	*0.32*	0.08	*0.10*
35-39	0.60	*0.69*	0.17	*0.23*
40-44	1.18	*1.18*	0.31	*0.42*
45-49	1.80	*1.77*	0.61	*0.71*
50-54	1.94	*1.96*	0.77	*1.00*
55-59	2.17	*2.24*	0.99	*1.17*
60-64	2.55	*2.57*	1.12	*1.39*
65-69	2.64	*2.76*	1.16	*1.56*
70-74	2.94	*2.98*	1.55	*1.65*
75-79	3.33	*3.45*	1.74	*1.88*
80-84	3.40	*2.98*	1.97	*2.07*
All ages (25-84)	1.72	*1.75*	0.76	*0.90*

Table IIB.2 HIP study group, participants. Screen-detected breast cancer at initial examination and subsequent rounds, by age at diagnosis.

Age at diagnosis	Initial examination			Rounds 2 - 4		
	Cases	Rate p.1000 examined		Cases	Rate p.1000 womanyears	
		HIP	*Simulated*		HIP	*Simulated*
40-49	16	1.72	*2.50*	15	0.76	*1.27*
50-54	12	2.61	*3.13*	19	1.56	*1.47*
55-59	15	3.88	*3.77*	20	1.87	*1.78*
> 60	12	4.95	*4.45*	23	2.67	*2.24*
All ages	55	2.73	*3.08*	77	1.49	*1.57*

Table IIB.3 HIP study group, participants. Screen-detected breast cancer in repeat rounds, by round.

Round	Cases	Rate p. 1000 examined	
		HIP	*Simulated*
2	32	1.85	*1.66*
3	18	1.05	*1.43*
4	27	1.57	*1.60*
Total	77	1.49	*1.57*

Table IIB.4 HIP study group, participants. Interval cancers, by round of the last screening and duration since the last screening.

Round	Duration (months)	Cases	Rate per 1000 woman-years	
			HIP	*Simulated*
1	<12	14	0.71	*1.05*
	>12	25	1.38	*0.89*
2-4	<12	32	0.75	*1.97*
2,3	>12	33	1.93	*1.82*
4	>12	68	1.95	*1.78*
Total		173	1.29	*1.42*

Table IIB.5 HIP study group, participants. Interval cancers, by age at diagnosis.

Age at diagnosis	Cases	Rate per 1000 woman-years	
		HIP	*Simulated*
40-49	51	1.29	*1.04*
50-54	38	1.17	*1.25*
55-59	35	1.19	*1.60*
> 60	49	1.50	*1.96*
Total	173	1.29	*1.42*

Table IIB.6 HIP study group, participants. Interval cancers by duration since last negative screening result.

Interval (months)	Cases	Rate per 1000 woman-years	
		HIP	*Simulated*
0- 5	9	0.28	*0.72*
6-12	38	1.19	*1.16*
12-23	44	1.58	*1.49*
24-35	34	1.67	*1.91*
36-47	27	1.94	*1.93*
> 48	21	2.00	*2.34*
Total	173	1.29	*1.42*

Table IIB.7 HIP control group. Breast cancer cases by age at diagnosis.

Age at diagnosis	Cases	Rate per 1000 woman-years	
		HIP	*Simulated*
40-49	98	1.62	*1.70*
50-54	109	2.21	*1.96*
55-59	97	2.15	*2.30*
> 60	139	2.63	*2.77*
Total	443	2.13	*2.13*

Table IIB.8 HIP control group. Breast cancer deaths, and case-mortality with underlying cause breast cancer, by age of entry.

Age at entry	Number of deaths	Case mortality	
		HIP	*Simulated*
40-44	39	0.561	*0.519*
45-49	43	0.536	*0.506*
50-54	54	0.482	*0.527*
55-59	43	0.555	*0.512*
60-64	33	0.507	*0.484*
Total	212	0.508	*0.512*

Table IIB.9 HIP study group, participants. Distribution of screen-detected breast cancer according to modality, at initial screening and subsequent rounds. Simulated (sim) percentages as predicted by MISCAN.

	Physical examination only			Mammography only			Both positive			Total
	HIP		*sim*	HIP		*sim*	HIP		*sim*	HIP
	N	%	%	N	%	%	N	%	%	N
Initial screen	24	44	*40.4*	21	38	*29.2*	10	18	*30.4*	55
Repeat rounds	35	45	*43.2*	23	30	*29.0*	19	25	*27.8*	77

Table IIB.10 HIP study, participants. Distribution of screen-detected breast cancers according to tumour diameter. Observed numbers and percentages adjusted for tumours with unknown diameter based on the survival experience of these tumours. Simulated (sim) percentages as predicted by MISCAN.

	≤ 2 cm.			> 2 cm.			HIP Total	
	HIP		*sim*	HIP		*sim*	known size	unknown size
	N	%	%	N	%	%	N	N
Initial screen	16	40	*52.8*	26	60	*47.2*	42	13
Repeat rounds	30	67	*69.4*	21	33	*30.6*	51	26
Interval cancers	56	43	*38.4*	72	57	*61.6*	128	45

Table IIB.11 HIP study. Difference in breast cancer deaths in study group and control group, for four different time-periods since entry, and case-mortality rates at the end of the time period.

	Time since entry into study (years)			
	0-4	5-6	7-9	10-14
Control group deaths	63	61	50	38
HIP case mortality	0.144	0.284	0.401	0.492
simulated case mortality	*0.148*	*0.265*	*0.398*	*0.512*
Study group deaths	39	42	42	42
HIP case mortality	0.090	0.191	0.294	0.403
simulated case mortality	*0.115*	*0.211*	*0.324*	*0.423*
% difference (HIP)	-38%	-31%	-16%	+11%
% difference (simulated)	*-22%*	*-18%*	*-16%*	*-12%*

III. A MODEL FOR BREAST CANCER SCREENING

Introduction

Policy recommendations for breast cancer screening primarily address the eligible age-group and the interval between subsequent screening tests. The choice of a policy should preferably be based on the balance between expected health effects and costs.

However, the relationship between the age-group and interval, and the health effects is far from straightforward. Even the results of experimental screening projects can give rise to very different recommendations. For example, in the Swedish randomized trials, there is not yet firm evidence for a mortality rate reduction for women below the age of 50. In this age-group, the proportion of breast cancers that had been detected by screening was substantially lower than in older age-groups (Tabar *et al.*, 1985; 1987). In Sweden itself, it was concluded that the screening interval had been too long, which resulted in a recommendation for annual screening in women between the ages 40 and 49 and biannual screening for older women (Tabar *et al.*, 1988). Other recommendations, however, state that, in view of this lack of evidence for a mortality rate reduction, screening should not start before age 50 except in an experimental setting (Gezondheidsraad, 1987; Van der Maas *et al.*, 1989; Forrest, 1986).

Some of the large number of factors that influence the relationship between screening policy and health effects are related to the natural history of breast cancer and the characteristics of the screening tests: the incidence of (pre-) clinical breast cancer, the duration of screen-detectable preclinical stages, the improvement in prognosis following early detection, and the sensitivity and specificity of mammography. Other factors determine the population coverage, for example, the attendance and the question of whether nonattenders constitute a high risk group. Nearly all of these factors are age-dependent.

In view of this complexity, mathematical models, in which these factors are included, are indispensable for making predictions about the health effects and costs of different screening policies. But having a model is not sufficient. To obtain valid predictions, the model assumptions should be tested against available data of existing screening projects.

The analysis reported here uses a model-based approach to evaluate screening programs, using the computer simulation package MISCAN (Habbema *et al.*, 1984; Habbema *et al.*, 1987). This approach has been used in cost-effectiveness analyses of screening for cervical cancer (Habbema *et al.*, 1987) and for breast cancer (Van der Maas *et al.*, 1989).

In this chapter, the breast cancer screening model will be presented and discussed. An earlier model, resulting from an analysis of the data from the HIP study (Chapter II), is not applicable for the current situation in view of the great improvements in mammographic technique since the time of the HIP study. The major differences between the HIP model and the current model will be discussed.

A detailed description of the model is presented in Appendix IIIA.

Material

The model-assumptions about preclinical disease and mammography were checked against the following results of the Dutch screening projects in Nijmegen and Utrecht (Verbeek *et al.*, 1984; Collette *et al.*, 1984): (1) detection rates by age and screening round; (2) incidence of interval cancers, by age, screening round and by time since last screening; (3) stage-distribution of screen-detected cancers, interval cancers, and unscreened cancers.

For Nijmegen, the results pertain to women in the birth-cohorts (1910-1939) invited for the first round. This wide age-range allowed for an age-specific comparison of model outcomes and screening results. For Utrecht, only results for birth-cohorts from 1911 to 1925 from the city of Utrecht were used (Table III.1).

Copies of the key data files from Nijmegen and Utrecht were kindly made available for the analysis and were used to derive the desired tabulations. The population files consist of records containing the participation pattern of all individual women, and the cancer files contain all breast cancers diagnosed in the study populations.

The population data files were also used to study the participation pattern. The attendance figures are clearly age-dependent, with older women having the lowest participation rate. In Utrecht, only women attending a screening were invited for the next screening round. In Nijmegen, where all women were (and are) invited for each round, it appeared that attendance decreased steadily in subsequent rounds, from 85% in round 1 to 67% in round 4 (ages < 65 years). Attendance rates are high in the participants of the preceding round and low in the nonparticipants. These findings were incorporated in the model.

The survival rates of clinically diagnosed breast cancers are based on tumor-size specific data from the Utrecht project (DOM, 1986), and on age-specific data from Sweden (Adami *et al.*, 1986). Model-assumptions about the effectiveness of screening in reducing breast cancer the mortality rate were tested against the results of the Kopparberg / Ostergötland trial (Tabar *et al.*, 1988).

Table III.1 Characteristics and results of the 2 breast cancer screening projects in the Netherlands (Verbeek *et al.*, 1984; Collette *et al.*, 1984; DOM, 1986; Collette *et al.*, 1988). Values refer to cohorts and rounds that were used in the present analysis.

	PROJECT	
	Utrecht (DOM I, city)	Nijmegen
Characteristics:		
Start year	1974	1975
Cohorts invited	1911-1925	1910-1939
age at first round	50-64	35-64
# Women invited (total)	20,600	23,000
Women invited in subsequent rounds	attenders only	all women
# rounds	5	4
Interval (years)	1, 1½, 2, 4	2, 2, 2
Screening tests	mammography, physical examination	mammography
Results:		
Attendance rate, round 1	72%	85%
Screen-detected round 1	108 (7.2*)	74 (3.8*)
breast cancer round 2-5	81 (2.0*)	120 (2.6*)
Interval cancers	77 (1.1**)	99 (0.9**)

* per 1000 women examined
** per 1000 woman-years

Methods

The MISCAN computer simulation package (Habbema *et al.*, 1984; Habbema *et al.*, 1987) was used for simulating the screening projects in Nijmegen and Utrecht according to the characteristics of these projects. The differences between the two projects in age-groups invited, screening intervals, and attendance patterns (see Table III.1) have been modeled explicitly. In the Utrecht variant of the model, screening by mammography and physical examination is simulated, and the sensitivity of the combined tests will be higher than in Nijmegen. Otherwise, and in particular with respect to the course of breast cancer and the sensitivity of mammography, the two variants of the model were kept identical.

The results of the simulations of the Nijmegen and Utrecht projects were compared with the observed data, using chi-square goodness of fit tests and, when appropriate, a *t*-test for equality of mean values for each table considered.

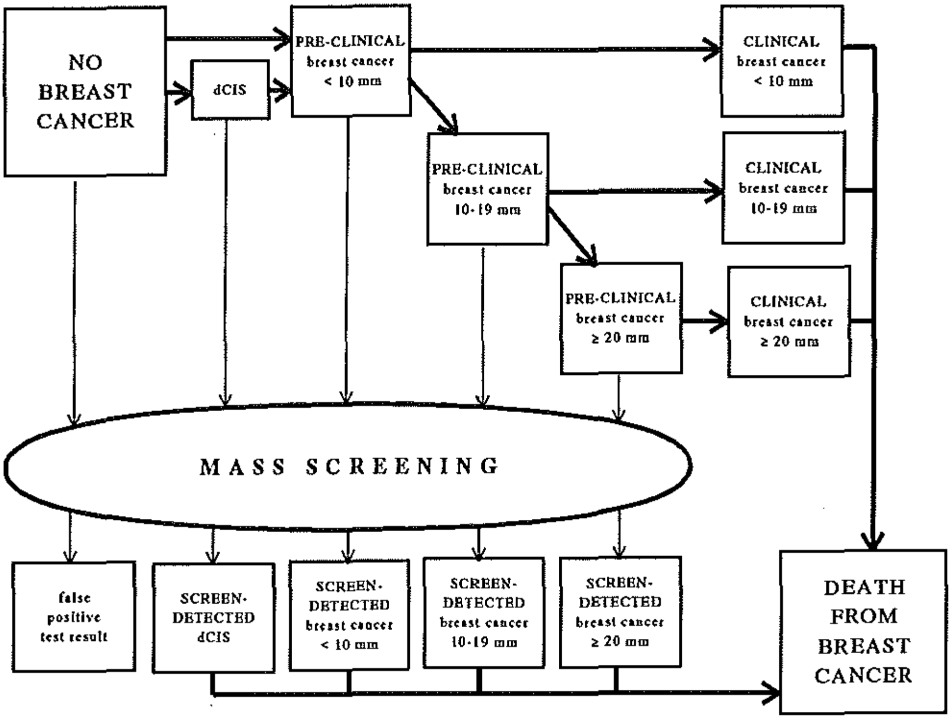
NO BREAST CANCER
dCIS
PRE-CLINICAL breast cancer < 10 mm
CLINICAL breast cancer < 10 mm
PRE-CLINICAL breast cancer 10-19 mm
CLINICAL breast cancer 10-19 mm
PRE-CLINICAL breast cancer ≥ 20 mm
CLINICAL breast cancer ≥ 20 mm
MASS SCREENING
false positive test result
SCREEN-DETECTED dCIS
SCREEN-DETECTED breast cancer < 10 mm
SCREEN-DETECTED breast cancer 10-19 mm
SCREEN-DETECTED breast cancer ≥ 20 mm
DEATH FROM BREAST CANCER

Figure IIIa Structure of the disease model for breast cancer, and the stages (tumor diameter categories) used in the model. The possible courses of the disease are indicated. The state death from other causes, which can be attained from all other states, is not shown.

We decided not to summarize the test results into one score for the overall goodness of fit of the model. There are many problems in combining results of statistical testing between tables from a screening project and between projects. Moreover, an overall lack of fit is not very informative; one has to go back to the individual tables to find the reason for the lack of fit. Therefore, the overall judgment of the adequacy of the fit is based on test outcomes for the separate tables. The tables in which a persistent lack of fit occurred are presented.

Model structure

The structure of the disease model is presented in Figure IIIa. The first state is NO BREAST CANCER. Women reside in this state until a transition occurs to one of the pre-clinical states that are detectable by screening (mammography or clinical examination). There is one pre-invasive state, intraductal carcinoma in situ (dCIS). The screen-detectable invasive state (pre-clinical breast cancer), is subdivided according to the diameter of the tumor: < 10 mm, 10 to 19 mm, and ≥ 20 mm.

The same subdivision is used for the symptomatic clinical states (clinically diagnosed breast cancer) and for the SCREEN-DETECTED BREAST CANCER states. The state FALSE-POSITIVE TEST RESULTS refers to women with a positive screening examination in whom no breast cancer is found at further assessment.

The two end states of the model are DEATH FROM BREAST CANCER and DEATH FROM OTHER CAUSES. The rate of transition to this state is governed by the Dutch lifetable, which is corrected for death from breast cancer (CBS, 1985).

The tumor diameter was given a central role in the stage distribution within the model because it determines the probability that breast cancer will be detected by screening; however is also affects the prognosis after treatment. In assessing screening policies, the use of the diameter-based staging will not only result in model-output concerning the number of cases that are detected early, but also concerning the stage-distribution and, thus, the prognosis of these cases. Additionally, it enables calculation of the impact of screening on the use of diagnostic and therapeutic procedures.

Having defined the structure of the model, the task of specifying its many parameters remains. Some parameter values can be calculated directly from available data. Examples are the preclinical incidence, the stage distribution and survival of clinically diagnosed cancers, and the attendance patterns in the screening projects. However, a number of important parameters cannot be obtained in this way; suitable values have to be obtained by experimentation with the model. These parameters are (1) the duration of screen-detectable invasive stages, (2) the sensitivity of the screening test, and (3) the improvement in prognosis following early detection.

In model experimentation, simulated results are compared with data from the Utrecht and Nijmegen projects to test the parameter values. An initial set of parameter values, which were partly based on the HIP model, resulted in many discrepancies between simulated and observed data. Next, the input parameters of the model were varied systematically. Eventually, a set of model specifications was obtained that gave an adequate overall fit of almost all screening results from the Nijmegen and Utrecht projects.

First, the parameter values that resulted in this fit will be presented. Then a comparison will be made of the simulated and the observed data.

Model parameters

Table III.2 summarizes the values of the key parameters of the model that resulted from the analysis.

Clinical breast cancer and survival

The clinical incidence is based on Dutch figures from 1977 to 1982 (CBS, 1985). The age-specific rates for more recent years are very similar, so there is no reason to assume important time or cohort trends for The Netherlands. The distribution of the tumor diameter for clinically diagnosed cancers was obtained directly from data on cancers diagnosed outside the screening programme in Utrecht and Nijmegen: 10% <10 mm, 22% between 10 and 19 mm, and 68% ≥20 mm. In the model, the distribution is independent from age.

Age- and (diameter)stage-specific survival data are available in the Utrecht project for a limited age-range. Long-term results are still lacking. Because of the rather small numbers, a parametric approach was used to determine the fraction cured and a probability distribution function for the survival time of cases that will die from breast cancer. The lognormal distribution parameters (mean and variance) were taken from an analysis of Swedish survival data (Rutqvist, 1985). Overall, approximately 58% of the breast cancer deaths in clinically diagnosed cases are assumed to occur within 5 years from diagnosis and 86% within 10 years. After 10 years, there is still a considerable excess in the mortality rate. The average fraction cured was estimated from the Utrecht data on clinically diagnosed cancers and varies with age according to the age-specific pattern in Sweden (Adami, *et al.*, 1986). The probability of survival (20-year follow-up, all stages combined) is highest (59%) around the age of 45 and declines with increasing age to 35% at the age of 80 (Table III.2).

The combination of model-assumptions on clinical incidence, stage distribution, and survival result in a good fit for the mortality rate for breast cancer in the Netherlands at all ages. When age-dependency of the survival is not taken into account, the simulated mortality for older women will be far too low: in the age groups 75 to 79, 80 to 84, and older than 85, the difference would be 5%, 17%, and 36%, respectively.

Table III.2 Key assumptions of the breast cancer screening model.

Parameter	Assumption	
Preclinical incidence	Based on Dutch clinical incidence, 1977-82	
Clinical stage distribution	% in stage	
< 10 mm.	10%	Independent
10-19 mm.	22%	from age
≥ 20 mm.	68%	
20-years survival of clinically diagnosed breast cancers (diagnosis at age 55)	Survival	
< 10 mm.	83%	Age-dependent
10-19 mm.	68%	(see text)
≥ 20 mm.	51%	
Duration of preclinical invasive stages	Average duration (years)	
Age 40	1.6	
Age 50	2.1	
Age 60	3.0	
Age 70	4.7	
Sensitivity of mammography	Sensitivity	
dCIS or < 10 mm.	70%	Independent
≥ 10 mm.	95%	from age
Impact of early detection: mortality reduction for screen-detected cancers	Reduction 52%	

Duration of preclinical stages

The average duration of the preclinical stage is proportional to the ratio between the detection rate in the initial screening round and the clinical incidence rate. This ratio clearly increases with age: in Sweden, from 1.9 in the age group from 40 to 49 years to 3.5 in age group from 60 to 69 years (Day *et al.*, 1988), and in Nijmegen from 2.1 to 3.1 for the same age groups.

When searching for appropriate model-assumptions about the relationship between the average duration and age, the model was tested against the age-specific detection rates and the age-specific incidence of interval cancers as observed in Nijmegen. A reasonable fit was obtained using the following assumptions and parameter values:

1. The average total length of the preclinical period, as given in Table III.2.
2. Ratios 4:6:7 for the average durations of the disease in the preclinical stages $<$ 10 mm : 10 to 19 mm : $\geq$ 20 mm.
3. The duration of the preclinical stages follows an exponential distribution. This distribution has been proven to give an adequate fit in the HIP model (see Chapter II), as in other model-based analyses (Walter and Day, 1983; Day and Walter, 1984).
4. For the stages dCIS and preclinical $<$ 10 mm, the sensitivity of mammography is 70%. For tumors $\geq$ 10 mm in diameter, the value is 95%.

The data from Utrecht and Nijmegen show that 12% to 15% of the cancers detected in the initial screen and 6% to 7% of those detected in the subsequent rounds were dCIS. It is (rather arbitrarily) assumed in the model that 5% of invasive breast cancers are preceded by a screen-detectable dCIS stage and that this stage lasts 7 years on average. Note that this is approximately three times as long as the preclinical invasive duration around ages 50 to 60 (see Table III.2), and thus, the proportion of dCIS in the prevalence of all preclinical stages will be approximately 15%.

It is assumed that all screen-detectable dCIS cases are progressive, *i.e.*, they eventually develop into an invasive form of breast cancer. This assumption cannot be justified on the basis of available data. One could just as readily assume that some fraction of dCIS cases are nonprogressive and still obtain a good fit by shortening the average duration. The 100% progression assumption is just made to complete the model. In application of the model to predict effects of different screening policies, this is one of the parameters for which a sensitivity analysis should be carried out. The proportion of dCIS in the cancers diagnosed outside mass screening is considered to be negligible and is assumed to be zero.

The age-specific incidence of the preclinical stage is obtained by shifting the clinical incidence rate to younger ages, by a number of years corresponding to the average (age-specific) total duration of the preclinical stages. This implies 100% progression of preclinical disease and excludes overdiagnosis by screening (except for inevitable cases in which women die from other independent causes in the period between screen-detection and clinical diagnosis without screening). This assumption is in agreement with findings from the randomized trials (Chu *et al.*, 1988; Tabar *et al.*, 1987).

The specificity of the screening test can be estimated directly from observed results of screening projects, for example *e.g.*, from Utrecht and Nijmegen. The model-assumptions used in cost-effectiveness calculations are given elsewhere(Van der Maas *et al.* 1989).

Improvement in prognosis following detection by screening

In the model, the (age-stage specific) survival of interval cancers and cancers diagnosed in nonparticipants is the same as the survival of clinically diagnosed cancers in the absence of screening. A decrease in the mortality rate in a screened population is a result of improvement in the prognosis of screen-detected cancers. This improvement has been estimated from the results of the 8-year Kopparberg/Ostergötland trial. When using the number of breast cancer deaths (153) among breast cancer cases in the control group in the trial for comparison with the study group, a correction has to be made for differences in size between both groups (Tabar *et al.*, 1988). After correction, 175 deaths would have been expected for the study group in the case of no screening. The total observed deaths in the study group is only 124 cases, 47 of which occurred in screen-detected cancers. If it is assumed that the difference of 51 deaths is due to the early detection and treatment, then the reduction in the breast cancer mortality rate for the screen-detected breast cancers is: 51 / (51+47) = 52%.

This 52% improvement in prognosis could not be explained entirely by the shift in tumor size category that results from early detection. The simulated decrease in the breast cancer mortality rate among screen-detected women is then only 40%. The additional improvement undoubtedly results from the tendency that also *within* a size category, screen-detected cases are detected relatively early (This especially applies to the highest tumor size category). Support for this hypothesis is given by the finding that within a size category, the observed proportion of node metastases is slightly lower in screen-detected cases (Tabar *et al.*, 1987). The same phenomenon can be observed in the Nijmegen and Utrecht data.

After combining the shifts within and between the screen-detected invasive stages, reduction factors have been obtained that express the reduction in risk of dying from breast cancer compared to this risk when the cancer had been diagnosed (possibly in a higher tumor size category) in the absence of screening. In the model, for stages <10 mm, 10 to 19 mm, and ≥20 mm, reduction factors of 0.75, 0.45, and 0.15, respectively, are assumed, and no risk of dying from breast cancer for screen-detected dCIS. Using these assumptions, the simulated reduction in the mortality rate after 5 to 8 years of follow-up observation is approximately 24% to 27%, which is within the range (23%-29%) reported for Kopparberg/Ostergötland (Tabar *et al.*, 1985; Tabar *et al.*, 1988).

Comparison of simulated and observed results

Figures IIIb-e compare the simulated results with the actual values from the projects in Utrecht and Nijmegen. The solid line in each figure represents the simulated values, and the dots give the actual values observed. The vertical lines through each point give the 95% confidence interval for the observed data using a normal approximation for the cases that are assumed to be governed by a Poisson distribution.

The model shows a good fit of the detection rate by screening round (Figure IIIb(1)) and by age-group (Figure IIIc) in Nijmegen, with the exception of the youngest age-group. The deviation in this age-group is statistically significant, although the number of cancers is relatively small. This discrepancy cannot be corrected by simply shortening the preclinical duration at younger ages, since this would cause a rise in the already too high interval cancer rate in this age-group, see Figure IIId. Assuming a lower sensitivity would have the same effect. It appears, however, that the simulated detection rate for the 35 to 44 age-group is approximately at the level that has been reported for rounds 5 and 6. (These data were not yet available for the present analysis.)

The same model-assumptions about the course of the disease and the sensitivity of mammography that gave good fitting results for Nijmegen, could not fit the detection rates analyzed from the Utrecht project, see Figure IIIb(2). We tried to obtain a better agreement between the model and the Utrecht data by varying the model-assumptions about the preclinical stage of breast cancer and the sensitivity of the screening tests. However, no assumptions could be found that resulted in a reasonable fit.

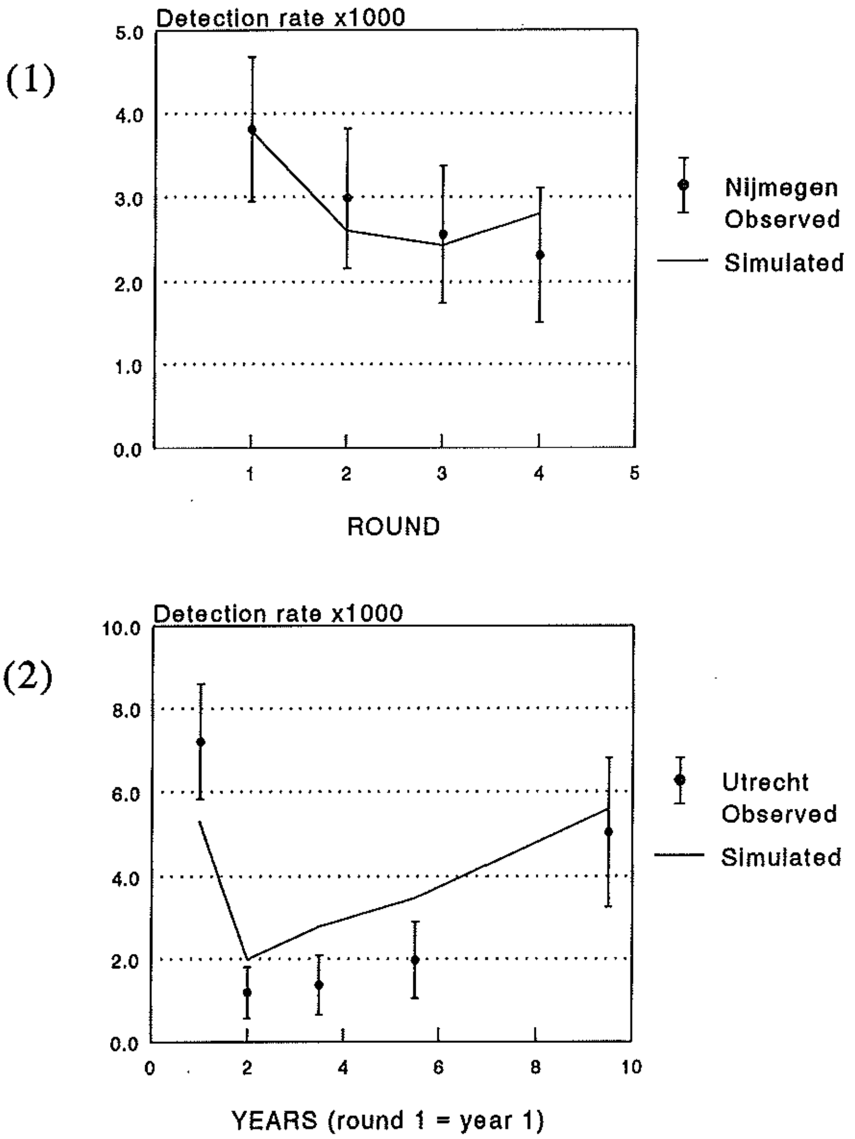

Figure IIIb Detection rates by screening round. Comparison of model results and the results from the projects in:
(1) Nijmegen (birth cohort 1910-1939); $p_{Chi}>0.5$, $p_T>0.5$.
(2) Utrecht (DOM I, city of Utrecht); $p_{Chi}<0.001$, $p_T>0.2$

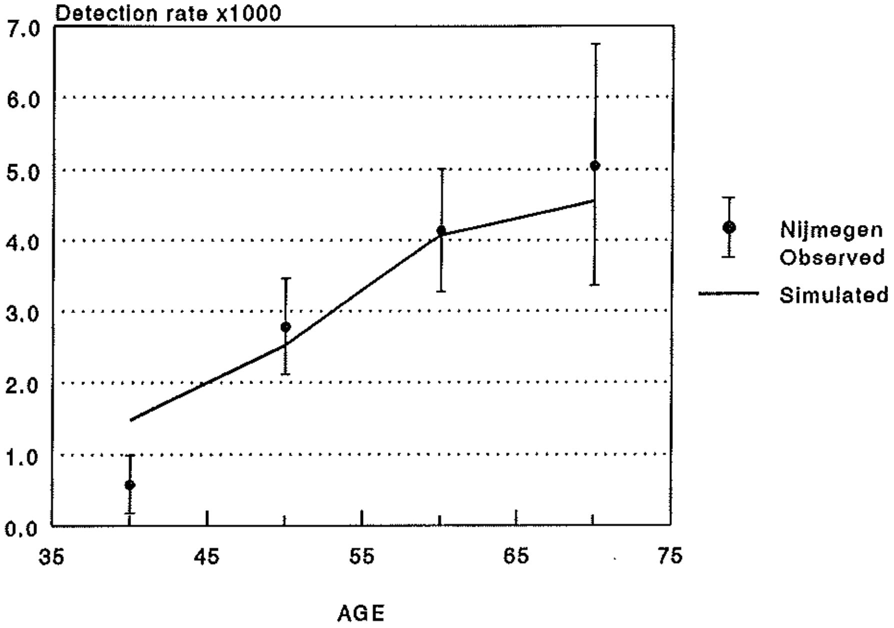

Figure IIIc Detection rate by age. Comparison of model results and the results from the Nijmegen project, birth cohort 1910-1939. $p_{Chi}>0.1$, $p_T>0.5$.

The lack of fit can to some extent be explained by the characteristics of the Utrecht project. For example, in the model, the result of a screening is dichotomous: either preclinical cancer is suspected, and the woman is referred for further diagnostic assessment, or not. In Utrecht, there were, however, two intermediate stages ("brief note to G.P." and "check-up message") that in a few cases led to the diagnosis of breast cancer some time later (Collette 1988). In addition, after the first round, only one view per breast was taken per screening, which may have affected the sensitivity. Both factors could have contributed to the low detection rates in rounds 2 through 4. However, these factors cannot explain such large differences between the model and observation in rounds 1 to 4.

The incidence rates for interval cancers show a good fit with the Nijmegen data, both in relation to age (Figure IIId) and to the length of time since the last screening (Figure IIIe(1)). This comparison in particular gives a good indication of the accuracy of the model-assumptions about the duration of the preclinical stages of the disease.

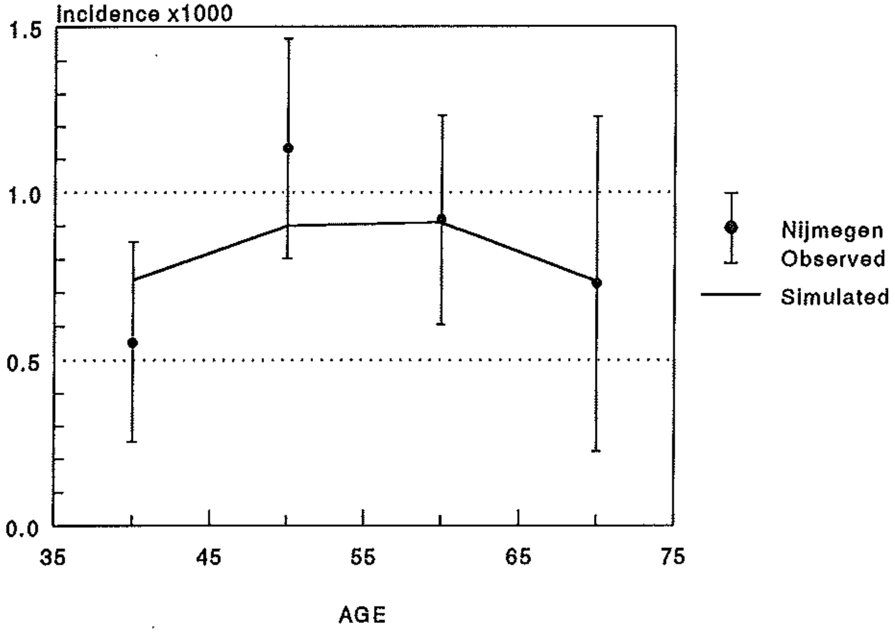

Figure IIId. Incidence rate of interval cancers, by age. Comparison of model results and the results from the Nijmegen project, birth cohort 1910-1939. $p_{Chi}>0.5$, $p_T>0.5$.

The same comparison for the Utrecht data shows again some deviations, see Figure IIIe(2). When these data are compared for each of the screening rounds separately (not shown), it appears that the simulated incidence of interval cancers is too high after rounds 1 and 2, and the simulated screen-detection rates in the immediately following rounds 2 and 3 are also too high (Figure IIIb(2)). For subsequent rounds, however, the too high simulated detection rate is counteracted by a too low simulated incidence of interval cancers.

Table III.3 presents a comparison between the simulated stage distribution of breast cancers and the distribution recorded in Nijmegen and Utrecht. The cancers are grouped according to the mode of detection. In general, there is either a fairly close correspondence between the model and the observed values, or else the simulated values lie between the values for Nijmegen and those for Utrecht. The simulated distribution for screen-detected cancers in the follow-up rounds is slightly too favorable in comparison with both projects.

The reverse is true for the interval cancers, especially for Utrecht, *i.e.*, there are too many tumors in the simulation with a diameter ≥ 20 mm.

Table III.3 Distribution of the breast cancers by stage of the disease at diagnosis.
Comparison of simulated and observed distributions for Nijmegen and Utrecht.
O = observed, S = simulated. True interval cancers: diagnosed before next screening round; Pseudo interval cancer: diagnosed after nonparticipation in a screening round.
Source: Data-files of the two screening projects.

Mode of detection		Total number	dCIS	≤ 9 mm	10-19 mm	≥ 20 mm
A. Nijmegen						
screen-detected						
round 1	O	74	12%	16%	32%	39%
	S		10%	25%	36%	29%
round 2-4	O	120	8%	25%	43%	24%
	S		8%	37%	39%	16%
true interval cancers						
round 1-4	O	99	9%	7%	28%	56%
	S		--	21%	24%	55%
unscreened						
	O	107	2%	21%	19%	59%
	S		--	12%	24%	64%
B. Utrecht (DOM I, city)						
screen-detected						
round 1	O	108	14%	25%	40%	21%
	S		10%	23%	37%	30%
round 1-4	O	81	6%	32%	46%	16%
	S		6%	39%	42%	13%
true interval cancers						
round 1-4	O	77	10%	14%	38%	38%
	S		--	24%	25%	51%
pseudo-interval						
	O	73	4%	11%	27%	58%
	S		--	16%	24%	60%
unscreened						
	O	172	2%	8%	33%	58%
	S		--	11%	23%	66%

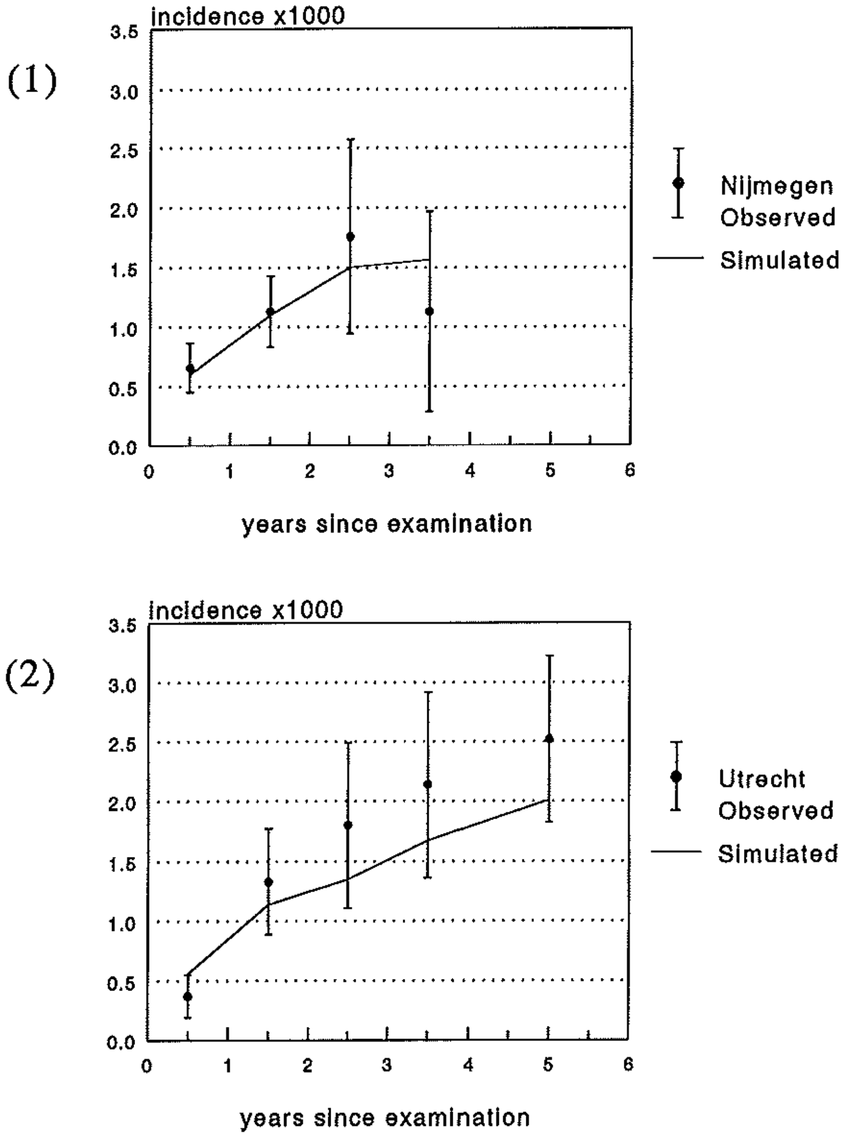

Figure IIIe. Incidence rate of interval cancers, by duration since last screening examination. Includes "pseudo" interval cancers detected after nonparticipation to a screening round. Comparison of model results and the results from the projects in:

(1) Nijmegen (birth cohort 1919-1939); $p_{Chi}>0.5$, $p_T>0.2$.

(2) Utrecht (DOM I, city of Utrecht); $p_{Chi}>0.1$, $p_T>0.1$.

The model-assumption that dCIS can only be detected by mass screening conflicts with the observed data for the interval cancers. It is, however, difficult to find an explanation for the observed percentage (approximately 10%) of dCIS in interval cancers, which is very high in comparison with the low percentage of dCIS in unscreened women. When dCIS is combined with the stage invasive cancer < 10 mm, then there is only a slight difference between the model and the observed data.

Discussion

The fit of the model to the breast cancer screening data is quite satisfactory, with one major and a few minor exceptions. The major exception is the lack of fit for the detection rates in the five screening rounds of the Utrecht project (see Figure IIIb(2)). Only data from the DOM I project in the city of Utrecht were included in the comparison. Results from the other projects in the city and the suburbs of Utrecht (Collette *et al.*, 1988) were not yet available. It is noteworthy, however, that at the first informal inspection, the outcomes of the model seem to be in much better agreement with the results reported from these other projects.

The lack of fit for the Utrecht project is not due to the simultaneous analysis of the data from Utrecht and Nijmegen. Independent parameter variation for Utrecht did not result in a better fit.

The improvement in prognosis is crucial in assessing the effects of screening. The case-control studies in Nijmegen and Utrecht estimate a reduction of 50% and 70% in the breast cancer mortality rate in screened women compared to unscreened women, respectively. (Verbeek *et al.*, 1984; Collette *et al.*, 1984). This is much higher than the 29% reduction from the Kopparberg-Ostergötland study in Sweden that was used in fitting the model. Indeed, a portion of the breast cancers in the Kopparberg-Ostergötland study group occur in unscreened women, but since the attendance rate is approximately 90%, this can only explain a few percent of the difference in mortality reduction.

From a methodologic point of view, the Swedish results should be preferred, because they stem from a randomized trial. In the Nijmegen and Utrecht case-control studies, part of the difference between screened and unscreened results might have resulted from selection effects.

Apart from the questions about differences between Sweden and the Netherlands regarding the quality of the screening or data collection, and the clear difference in attendance levels, the screening situation in both countries is similar in many respects. For example, the Kopparberg/Ostergötland policy closely resembles the one used in Nijmegen, and the proportion of cancers below 20 mm. in diameter is 43% in Sweden (Tabar *et al.*, 1987) and 42% in the Netherlands (Table III.3).

More important, the results of the project in the United Kingdom (UK Trial of Early Detection of Breast Cancer Group, 1988) and the trial in Malmö (Habbema *et al.*, 1986), published after completion of the analysis reported here, both show more modest reductions in the mortality rate, even less than the Kopparberg/Ostergotland study. This brings the estimate used in the model to the comfortable position of being an intermediate estimate.

A number of presuppositions are made in the model. The clinical stage distribution (tumor diameter) is assumed to be constant for all ages. This is probably not correct, but the practical implications are minor since the impact on prognosis is already incorporated in the age-dependent survival.

From a prognostic point of view, refinement of the present stage distribution into lymph node/tumor diameter stages is attractive. However, the model would become more complicated, and finding appropriate parameters for the preclinical stages would become even more tedious. The current approach in which the improvement in prognosis is corrected for shifts within diameter stages is more feasible.

The sensitivity of the screening test depends on the tumor diameter but is assumed not to depend on the age of the woman. It is difficult to distinguish between the concepts of sensitivity and duration of screen-detectable stage. Model assumptions of a shorter duration of preclinical disease at younger ages can to some extent be exchanged for assumptions stating that the sensitivity is lower in these age groups.

Another important issue is the effect of screening for women under 50 years of age. Thus far, the results for younger women are inconclusive (Tabar *et al.*, 1988; Chu *et al.*, 1988; Habbema *et al.*, 1986; Andersson *et al.*, 1988). In the model, the improvement in prognosis in screen-detected women is independent of age. However, since the preclinical duration is shorter at younger ages, and the survival of clinically diagnosed cancers is highest between the ages of 40 and 50, the model will predict a lower reduction (%) in mortality below the age of 50 than for older women. Given the ambiguity of current empirical results for effectiveness below the age of 50, the model should be used with care in these age groups. The lower incidence below age 50 further detracts from the desirability of screening in these ages.

In the model, there is no difference in the risk of breast cancer between participants and nonparticipants. In the HIP study, the risk in nonparticipants was estimated to be lower than average. In the Netherlands, a more indirect comparison between Nijmegen and the nearby city of Arnhem where no mass screening took place led to the conclusion that there was no clear indication for a difference in risk (Verbeek *et al.*, 1984). The same can be concluded from the Utrecht project by comparing the incidence in nonattenders with the incidence in the years before screening started.

The assumptions made for dCIS in the model are highly speculative, but are compatible with the findings in screen-detected cancers. The assumption of no dCIS among clinically diagnosed breast cancers is an oversimplification, especially for interval cancers. The importance of dCIS may well increase with additional improvement in screening techniques. For example, figures from Nijmegen for screening rounds 5 and 6 show a percentage of approximately 19% dCIS in screen-detected breast cancers.

Comparison with the analysis of HIP study

The MISCAN approach was also used to analyze the results of the HIP study (see Chapter II). However, the model consisted of two tumor diameter classes only, since tumors below 1 cm. in diameter were not screen-detectable at that time. No age-dependency was assumed for the average duration. Although the data suggested some age-dependency, the number of screen-detected cases in the HIP study did not allow for conclusive estimates. The average duration of the preclinical screen-detectable stage derived in the HIP analysis was 2.2 years (age group 40-70 years). This is the same as the average total duration of the dwelling time for tumors in the stages of 10 to 19 mm and ≥20 mm in the present model for women aged 55. The estimated sensitivity of mammography for the HIP model was only 50% (Chapter II) which is much lower than the 95% in the present model.

These comparisons show the potential power of modeling: the parameter values for the invariant part of the natural history of preclinical breast cancer are indeed the same, whereas the increase in the sensitivity reflects the improvement in mammography. Taking the obvious differences between HIP and Nijmegen into account, the model shows that there is a good correspondence between the screening data from these studies. Exploratory simulations of the projects in the UK and Sweden seem to indicate the promising prospect of fitting these studies as well.

Comparison with results from other models

The findings about the duration of preclinical disease and the sensitivity of screening can be compared with results from other modeling approaches. A review of these approaches is given by Prorok (Prorok, 1988).

The statistical model of Day and Walter (1984) has been applied to the results of several screening projects (Day *et al.*, 1988; Walter and Day, 1983; Verbeek *et al.*, 1988). In its original version, this model consists of a single preclinical stage, for which a probability distribution function for the dwelling time is specified. The parameters of this function, the sensitivity of the screening test, and the clinical incidence of breast cancer can be estimated from screening data. In some applications, the clinical incidence is fixed to a "known" value, *e.g.*, the incidence in the control group when a randomized study is analyzed. The model is confined to the detection of breast cancer; the effects on mortality rate are not taken into account. One advantage of this model is that it is relatively simple to obtain an approximate confidence interval for the parameter values.

This model was applied to data from Utrecht by Day *et al.* (1988). The study reports a good fit of the model (chi-square of 7.2 and 7 degrees of freedom [df]), when assuming a sensitivity of 99% and a mean duration of 2.8 years. It is not indicated exactly what data from Utrecht were used, but it is clearly a less detailed subset of the data that we used for testing model assumptions. When we simulate this simplified model with a mean duration of 2.8 years and 99% sensitivity and test the outcomes against the screen-detected cases in the five screening rounds of the Utrecht data (DOM I project, city of Utrecht), the chi-square rises to 16.0. When the interval cancers are also added to the analysis, the chi-square value becomes 50.9 for 22 df, indicating a very bad fit of the model. The largest deviations occur in rounds 3 and 4 (screen-detected and interval cases), so probably these have not been included in the analysis of Day *et al.*. It is important to note that this simpler model also fails to give a good description of the results from Utrecht.

An adapted version of the Day and Walter model was applied to the Nijmegen data (Verbeek *et al.*, 1988). In general, the estimated parameters are comparable to the values found in the MISCAN approach presented here, especially regarding age-dependency of the estimated duration of the preclinical stage. However, the reported average duration is somewhat shorter, *e.g.*, 2½ years in the 50 to 64 age group (compare Table III.2). This difference may be due to the fact that the sensitivity in the current model varies within the preclinical stage, which will result in an increase of the average duration.

Other model-based evaluations of breast cancer screening have been published (Eddy, 1980; Shwartz, 1978; Eddy *et al.*, 1988). However, in these evaluations not much attention is devoted to the subject of fitting the model to empirical data from large screening studies. The model of Shwartz is based on a biological formulation of tumor growth; parameter estimates for this model cannot be compared directly with our approach.

Adaptation of the model for cost-effectiveness calculations

The model presented here has subsequently been used in predicting the effects and costs of different screening policies for breast cancer. For this purpose, some adaptations and extensions are needed, but the core of the model, *i.e.* model-assumptions about preclinical disease, sensitivity of screening, and improvement in prognosis remains unchanged.

The adaptations mainly refer to the population studied: instead of specifying the Utrecht and Nijmegen populations, now the entire Dutch female population is considered, taking the expected dynamics in the forthcoming decades into account. During this period, the average attendance is assumed to remain constant at 65% in the age group of 50-69 years. The attendance pattern is modeled according to the experience in Nijmegen.

The clinical part of the model was extended, by linking schedules for diagnostic and treatment procedures to the screen-detected and the clinical states and to the state "false positive test result". Treatment of advanced disease is linked to the state "death from breast cancer". This extension is useful in determining the change in demand for diagnostic and therapeutic facilities following the introduction of mass screening (De Koning *et al.*, 1990).

Finally, cost estimates for all activities related to screening, diagnosis, and treatment have been linked to the model. This enabled calculation of costs and savings of different screening policies. The main conclusion from these calculations is that in terms of costs per life-year gained, biannual mammographic screening of women older than 50 years of age compares favorably with other health care facilities (Van der Maas *et al.*, 1989).

Appendix IIIA.

DETAILED OVERVIEW OF THE MODEL SPECIFICATIONS.

This appendix gives the full parameter quantification of the breast cancer screening model. Details on the simulation method can be found in the description of the MISCAN model (Habbema *et al.*, 1984).

Demography
Women die from other causes according to the Dutch lifetable for 1982-83, in which breast cancer deaths have been excluded.

AGE	Proportion dead
35	0.0149
40	0.0204
45	0.0280
50	0.0389
55	0.0558
60	0.0829
65	0.1273
70	0.1996
75	0.3128
80	0.4759

Risk groups
The female population is assumed to be homogeneous with respect to risk of breast cancer. Apart from the obvious factor age, no other risk factors are taken into account. A consequence of this assumption is that participation at screening is independent from risk.

Disease model

Note: the description in this section applies to the situation in which there is no screening.

Stages:
The stages involved in the disease process are represented in Figure IIIa. In addition, transitions to the stage DEATH FROM OTHER CAUSES are possible from all stages except DEATH FROM BREAST CANCER. These transitions are governed by the lifetable (see I).

Preclinical incidence.
The preclinical incidence is described by a piecewise linear probability distribution function, with a constant probability density within 5-year age-

groups. The preclinical incidence is estimated from the reported clinical incidence data. It is shifted to younger ages according to the assumptions about the transitions and durations in the preclinical stages. The relation between clinical incidence I_i in each of the N age-groups i and the preclinical incidence P_j in preceding age-groups j is given by:

$$I_i = \sum_{j=1}^{i} D(i,j) P_j \ , \quad i = 1,..,N$$

A deterministic auxiliary model was used to calculate the matrix D(i,j) representing the probability of clinical diagnosis for women who entered the first preclinical stage in age-group j. The P_j values are then calculated by solving the set of N equations.
The preclinical incidence rates are:

Age group	Preclinical incidence (per 100,000)
35-39	94
40-44	163
45-49	216
50-54	221
55-59	264
60-64	297
65-69	345
70-74	421
75-79	472
80-84	464

Preclinical course.

From the preclinical incidence, 5% pass through the preclinical screen-detectable stage dCIS. The dwelling time in this stage follows an exponential distribution with mean 7 years. Transitions between stages are described by transition probabilities p_T, the dwelling time distribution in a stage is conditional on the next stage and follows an exponential distribution. The mean durations m(a) are age-dependent according to the formula (m_{50} is the mean duration at age 50):

$$m(a) = \frac{0.3\ m_{50}}{(1.0 - 0.9a)^2} \ , \quad a = \frac{age}{100}$$

From	To	p_T	m_{50}
PRECLIN <10 mm	PRECLIN 10-19 mm	0.9	0.58
PRECLIN <10 mm	CLIN <10 mm	0.1	0.58
PRECLIN 10-19 mm	PRECLIN ≥20 mm	0.75	0.85
PRECLIN 10-19 mm	CLIN 10-19 mm	0.25	0.85
PRECLIN ≥20 mm	CLIN ≥20 mm	1.0	1.00

Survival.

Survival is described by a fraction cured and a survival time distribution for women who are at risk of dying from breast cancer. Values for the mean and variance of the log-normal survival time distribution are adopted from an analysis reported by the Swedish Cancer registry (Rutqvist, 1985). The fraction cured in each clinical stage, as estimated from survival data from Utrecht, is:

Stage	Fraction cured
CLIN <10 mm	0.757
CLIN 10-19 mm	0.702
CLIN ≥20 mm	0.407

Mass screening: ages, recall scheme, participation

Nijmegen variant

The age distribution of the population invited for the first screening round is specified by giving the relative size of the (5-year) age-groups involved. Total number invited: 23,000.

Age distribution and attendance, Round 1

Age group	Relative size	Attendance
35-39	0.119	0.875
40-44	0.171	0.872
45-49	0.172	0.870
50-54	0.179	0.852
55-59	0.169	0.852
60-64	0.160	0.814
65+	0.040	0.708

Recall scheme: all women in the original cohort are invited for rounds 2-4. The interval between rounds is 2 years. The attendance in a round was found to differ especially between participants and non-participants in the preceding round.

Attendance, rounds 2-4.

Round	Interval years	Attendance participants	Attendance nonparticipants
2	2	0.853	0.155
3	2	0.807	0.192
4	2	0.820	0.191

Screening test: mammography.

Utrecht variant. Total number invited: 20,600.

Age distribution and attendance, round 1.

Age group	Relative size	Attendance
50-54	0.360	0.762
55-59	0.323	0.726
60-64	0.317	0.650

Recall scheme: only participants at the preceding round are invited for a new round.

Attendance, rounds 2-4.

Round	Interval (years)	Attendance (participants)
2	1.0	0.808
3	1.5	0.860
4	2.0	0.844
5	4.0	0.714

Screening tests: mammography and physical examination.

Sensitivity

The sensitivity of the screening tests is defined for each "true" stage of the disease. Mammography and physical examination are assumed to be independent, i.e. the overall sensitivity in stage 10-19 mm. is 0.95 in Nijmegen and 0.975 in Utrecht.

State	Sensitivity	
	Mammography	Physical examination
dCIS	0.7	0.0
< 10 mm.	0.7	0.0
10-19 mm.	0.95	0.5
≥ 20 mm.	0.95	0.7

Improvement in prognosis

The improvement in prognosis resulting from early detection by screening is highest in the early stages.

Stage	Reduction in lethality from breast cancer.
dCIS	1.0
< 10 mm.	0.75
10-19 mm.	0.45
≥ 20 mm.	0.15

Testing the goodness of fit.

Simulated and observed are arranged in a number of tables, for example the tables for detection rate by round which correspond to Figures IIIb(1) and IIIb(2).

In comparing the simulated rates s and observed rates r, which are based on q and p cases, respectively, allowance should be made for the variance in the simulated rate which is inherent to the method of microsimulation (Habbema *et al.*, 1987). Each rate is tested using the formula:

$$Z = \frac{r-s}{\sqrt{r^2/p + s^2/q}}$$

When simulating a sufficiently large number of life-histories, q> >p and var(s) < < var(r), and total variance is dominated by the variance of the observed data. For each table, a chi-square value was calculated by adding the sum of squares of the Z values in the categories in the table, with the number of degrees of freedom equal to the number of categories. A t-test for the mean value of the rates in a table is directly based on the Z value.

In figures IIIb-e, the vertical lines denote confidence intervals of the observed rates. Variability of simulated rates is not displayed in the figures.

IV. EPIDEMIOLOGICAL EVIDENCE FOR AGE-DEPENDENT REGRESSION OF PRE-INVASIVE CERVICAL CANCER

Introduction

Regression of pre-invasive cervical cancer is still a controversial issue. The possibility of regression, which may not occur at the same rate at all ages, is important for decision making about screening policies and about management of these lesions. A high regression rate would mean that many screen-detected lesions are diagnosed and treated unnecessarily. This would influence the balance between favourable and adverse effects of screening.

In the analysis of the screening data from the British Columbia cohort study, a considerable amount of regression was found. The regression is concentrated in lesions detected at younger ages (Boyes *et al.*, 1982). This finding was based on a comparison of the estimated cumulative incidence of pre-invasive lesions with the prevalence of these lesions and with the cumulative incidence of clinically diagnosed invasive cervical cancer. However, no explicit hypotheses about regression were tested. A model based approach would allow for using the detailed age-specific data from the cohort study in testing the assumptions. Earlier modelling efforts, e.g. by Coppleson and Brown (1975) also pointed at the existence of regression, but used a combination of data from widely different sources.

For breast cancer, Day and Walter (1984) proposed a simplified model that can be used in analyzing data from screening programs. In this paper we propose a simplified model for cervical cancer, and use the model to test hypotheses about regression against data from the screening program in British Columbia.

A model for cervical cancer is more complicated than for breast cancer, because of the long duration of pre-invasive stages and the possibility of regression. The proposed model contains the essentials of screening for cervical cancer: incidence of pre-invasive lesions, duration, progression and possible regression of these lesions, and sensitivity of the Pap smear.

Three hypothesis about regression of pre-invasive lesions are tested: (A) no regression, (B) a constant rate of regression at all ages, and (C) different regression rates for lesions that occur in young and middle-aged women, respectively.

Material and methods

The data used in testing the hypotheses are all derived from the report of the British Columbia cohort study (Boyes *et al.*, 1982). This study analyzed the records of the screening program in British Columbia in the period 1949-1969, for two cohorts of women, born in 1914-18 and 1929-33, respectively. A full overview of the data is presented in Appendix IVA, Table IVA.1.

The age-specific incidence of clinical cervical cancer is based on data for the total female population between ages 20-24 and 60-64 in British Columbia in 1955-1957. Cervical cancer screening started already around 1950 in British Columbia, but only on a very limited scale. The impact on clinical incidence in the first years of screening has presumably been negligible. Thus, the incidence in 1955-57 is considered to represent the situation without screening. We converted the clinical incidence into an incidence rate for the female population at risk of cervical cancer by correcting for the cumulative hysterectomy rates recorded in the two cohort study populations.

A distinct advantage of the data from British Columbia is that age-specific rates are available over a broad age-range, both for clinically diagnosed cancers and for screen-detected lesions. The two cohorts together span the age-range 20-54 in the study period. In the overlapping age-group 35-39, results from the younger cohort are used because of the larger numbers in this group. It is assumed that the cross-sectional clinical incidence data and the longitudinal screening data are comparable. Differences between these data sets could be caused by cohort effects, but cohort differences were not found in the cohort study (Boyes *et al.*, 1982).

Regression is supposed to occur only in pre-invasive lesions. The observed detection rates of pre-invasive lesions (dysplasia and carcinoma in situ) of a first smear are included in estimating and testing the assumptions about regression (see Table IVA.1). In the results of the second and subsequent smears the pre-invasive and early invasive lesions are pooled, because of the small numbers in the latter category. On basis of the tables in the report of the cohort study, these results are classified by age at midpoint between the first and second smear, and by the time since the preceding Pap-smear.

The clinical incidence rates in the unscreened parts of the cohorts are also used in testing the model. An important reason for including these data in the analysis is that they contain information about the difference in potential risk for cervical cancer between participants and non-participants in screening.

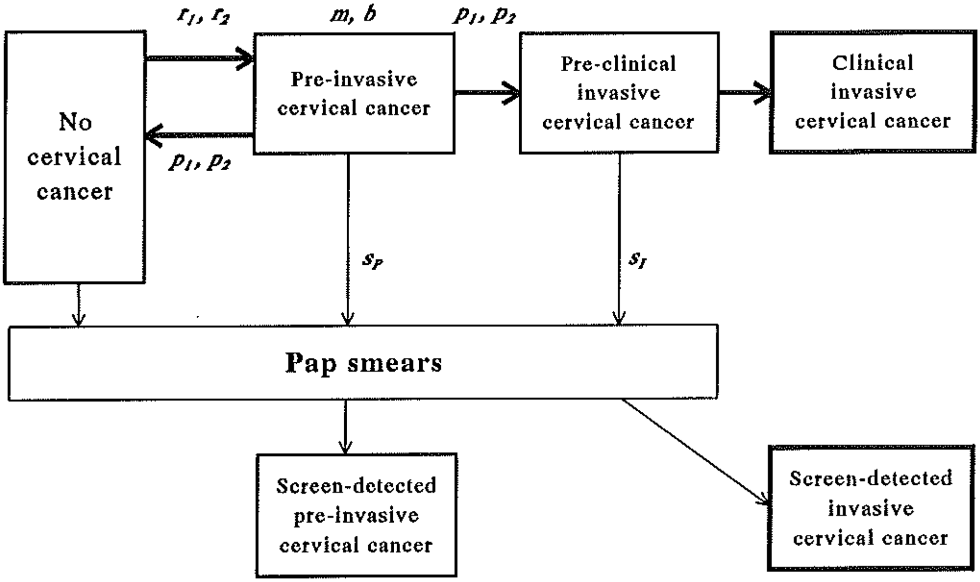
r_1, r_2
m, b
p_1, p_2
No cervical cancer
Pre-invasive cervical cancer
Pre-clinical invasive cervical cancer
Clinical invasive cervical cancer
p_1, p_2
s_P
s_I
Pap smears
Screen-detected pre-invasive cervical cancer
Screen-detected invasive cervical cancer

Figure IV.a Schematic representation of the model.

The model

The model consists of 5 states: NO CERVICAL CANCER, PRE-INVASIVE CERVICAL CANCER, PRE-CLINICAL INVASIVE CERVICAL CANCER, SCREEN-DETECTED LESIONS, and CLINICAL CERVICAL CANCER, see Figure IV.a. Death from other causes and hysterectomies for other reasons than cervical cancer are treated as independent exogenous factors that have no influence on incidence rates in the "at risk" population or on screen-detection rates. Survival of screen-detected and clinical cervical cancer and death from cervical cancer are not considered in the model since they do not relate to the problem of regression.

The model is stochastic, i.e., the disease process is described in probabilistic terms. The transitions between stages are described by probabilities, and the dwelling time within the stage PRE-INVASIVE CERVICAL CANCER is governed by a probability distribution.

The onset of pre-invasive screen-detectable stages corresponding with dysplasia and Carcinoma in Situ is reflected in the transition from NO CERVICAL CANCER to PRE-INVASIVE CERVICAL CANCER. This onset is assumed to be age-dependent and starts at age a_0. Women who participate in screening are assumed to have a relative risk rr in comparison with the total population.

From PRE-INVASIVE CERVICAL CANCER, the lesion may regress spontaneously (i.e., return to NO CERVICAL CANCER) or it may progress into PRE-CLINICAL INVASIVE CERVICAL CANCER. The proportion of cases in which progression occurs is age-dependent. The state PRE-CLINICAL INVASIVE CERVICAL CANCER consists of micro-invasive lesions and occult tumours. The duration in this state, i.e., the time between invasion and clinical diagnosis of cervical cancer, is assumed to be the same in all women. All invasive lesions are assumed to be progressive. This chain of probabilities will result in the (age-dependent) clinical incidence rate.

The lesions in stages PRE-INVASIVE CERVICAL CANCER and PRE-CLINICAL INVASIVE CANCER can be detected when a Pap smear is made. The sensitivity of this screening test is s_P for pre-invasive and s_I for invasive lesions, respectively. The detection rates for women who have a Pap smear will depend on their age, on the rank of the smear, and on the time since the preceding smear. The formulae for the clinical incidence rate and the detection rates at first, second, and subsequent Pap-smears are given in Appendix IVB.

In order to make the model more parsimonious, a number of simplifications are made. First, both the onset rate of pre-invasive lesions and the proportion of progressive lesions are assumed to have two levels. Onset rate r_1 and proportion progression p_1 apply to younger women (between ages a_0 and a_1), and levels r_2 and p_2 apply to women of age a_1 and older.

Second, the pre-invasive duration is described by a Weibull probability distribution which has two parameters: mean duration m and shape (or concentration parameter) b. The Weibull distribution is a generalisation of the exponential distribution. The additional concentration parameter allows for changing the variance in the duration independently from the mean, higher values indicating less variability.

Third, the duration of the preclinical invasive stage and the sensitivity in this stage are assessed directly from the clinical incidence rates and detection rates of the first smear, see Table IVA.4. (Note that these detection rates are not included in testing the model.) The ratio of the two rates, shown in the last column of the table, represents an estimate for the product of sensitivity and duration (approximately 3.6 years). The sensitivity for invasive lesions was fixed at s_I = 0.90, and the duration was set a 4 years. A last simplification was used in calculating the detection rates at second and subsequent smears, see Appendix IVB for details.

In analyzing the data, three models A, B, and C are compared. In model A, it is assumed that all pre-invasive cases progress (p_1=p_2=1.0). In models B and C only a certain proportion of pre-invasive lesions will progress to invasive cancer. In model B, this proportion is independent from age ($p_1=p_2$). In model C, the proportion of progressive lesions differs between young and middle-aged women ($p_1 \neq p_2$).

Model A has 8 parameters: the relative risk (*rr*), the ages and rates of the incidence of pre-invasive lesions (a_0 , a_1 , r_1 , r_2), the duration of pre-invasive lesions (*m*, *b*), and the sensitivity of the Pap smear (s_P), that are to estimated. By including the progression parameters (p_1 and p_2), model C has 10 parameters, while model B has only 9 independent parameters because the progression parameters are equal.

Estimation procedure

The cohort study data on clinical incidence (8 age-categories after combining 20-24 and 25-29 because of very small numbers), detection rates at the first smear (7 categories), detection rates at the second and at subsequent smears (2 x 4 x 3 categories), and the clinical incidence in unscreened parts of the cohorts (6 categories), are all arranged into a single table (see Table IVA.1) with 45 entries.

For each set of parameter values, expected number of cases according to the model are calculated. Both a Pearson Chi-square test statistic for the goodness of fit and the Likelihood are calculated from the expected and observed numbers of cases. For each model, best-fitting parameter estimates are obtained by maximization of the likelihood. The Likelihood-Ratio test is used to compare the models A, B, and C, and also in finding one- and two-dimensional confidence regions for the parameters. More specific details on the estimation and testing procedures are given in Appendix IVB.

Results

The best fitting parameter values for the three models (A, B and C) are presented in Table IV.1. The goodness-of-fit test for the assumption that all pre-invasive lesions will progress to become invasive cancers (model A) shows that it is not possible to fit the data from British Columbia with this assumption.

Especially the results of second and subsequent smears show large discrepancies between observed data and the model (See Appendix IVA, Table IVA.1). The clinical incidence and the detection rate of the first screen are fitted fairly well. This obviously requires quite surprising parameter estimates: a very long average duration (33 years) of pre-invasive lesions, coupled with an incidence rate of 1.3×10^{-3} between age 15 and 33 which is much higher than the clinical incidence rate. After age 33, few pre-invasive lesions start developing. Because of the low estimate (66%) for the sensitivity of the Pap smear, the model gives too high detection rates for smears made within a short interval after the preceding smear. And the relatively low incidence rate of new lesions results in much too low detection rates for smears made after an interval of more than 3 years since the first or second smear.

Table IV.1 Parameter estimates for pre-invasive lesions of cervical cancer, and goodness of fit of the three models.

Parameter	Model		
	A	B	C
	No regression	Regression, constant	Regression, age-dependent
Incidence of pre-invasive lesions (rates $\times 10^3$ woman years):			
start at age a_0:	15	15	18
change at age a_1:	33	33	34
incidence rate r_1 (before age a_1)	1.31	1.46	2.11
incidence rate r_2 (after age a_1)	0.16	0.54	1.06
relative risk rr of participants	0.80	0.75	0.74
Duration and progression of pre-invasive lesions:			
mean duration m (years)	33.3	21.5	11.8
shape of distribution b	2.06	2.37	1.58
progression p_1 (before age a_1)	1.00	0.47	0.16
progression p_2 (after age a_1)	1.00	0.47	0.60
Pap smear:			
sensitivity s_p	0.66	0.70	0.80
GOODNESS OF FIT:			
p-value	0.0001	0.005	0.7

With a shape parameter of 2.1 the variability in duration is rather low, and only 24% of the pre-invasive lesions will have a duration of less than 20 years. In other words: although no explicit regression is assumed in model A, this very long duration, with a considerable amount of (very) slow-progressing lesions, can be interpreted as a compensating mechanism.

The assumption of an equal proportion regression at all ages (model B) results in an estimate of 53% for the proportion regression among pre-invasive lesions, see Table IV.1. The estimates for the onset after age 33, the mean duration of pre-invasive lesions, and the sensitivity of the Pap smear differ considerably from the case with no regression (model A). Although model B gives a statistically significant ($p<0.0001$) better fit than model A, the goodness of fit test against the cohort study data still yields a p-value smaller than 0.01. The same discrepancies with observed data found in model A exist in model B; they are only less extreme.

The difference in log-likelihood between the model (A) with no regression and the model (C) with age-dependent regression indicates that a clearly significant ($p<<0.0001$) improvement is brought about by adding age-dependent regression. Moreover, model C gives a good fit of the observed data from British Columbia. Between age 18 and 34, the incidence of pre-invasive lesions is high, and the estimated proportion of regression among these lesions is 84%. The proportion regression over age 34 is 40%. From all lesions developing before age 65, an average of 62% is regressive. Estimates for the other parameters show considerable differences compared to the case with no regression (model A, see Table IV.1). In women older than 34 years, there is a substantial onset rate of new pre-invasive lesions. The estimates for the duration of pre-invasive lesions imply that the large majority (85%) of all new progressive lesions will turn into invasive lesions within 20 years. In combination with the higher sensitivity (0.80) of the Pap smear, these changes lead to a considerable improvement in the fit of the detection rate of second and subsequent smears. The relative risk of participants is 0.74, indicating that unscreened women constitute a high risk group.

For model (C), assumptions about average duration, sensitivity, progression rate, and the shape have been varied to find 95% confidence limits for these parameters, see Table IV.2. The mean duration of pre-invasive lesions is between 9.8 and 14.4 years. Mean durations of less than 9.8 years result in clinical incidence rates becoming too high at older ages, durations longer than 14.4 years will conversely result in too high detection rates at older age. The range for the shape parameter of the distribution means that the standard deviation is between 5.9 and 12.1 years. The sensitivity of the Pap smear for pre-invasive lesions is between 76% and 85%. Other values will especially deteriorate the fits of detection rates by interval at second and subsequent smears. The range for progression in lesions starting before age 34 is 8%-24%, indicating that at most a quarter of these lesions will become invasive. The lower bound is imposed by the clinical incidence at young ages. The upper bound is imposed by an overall tendency in all detection rates to become too low if progression is over 30%. The range for the proportion progression over age 34 is rather wide (42%-88%). Since 100% progression is not included in the range, the model supports the hypothesis that regression also occurs in lesions that develop at higher ages.

We also considered alternative values for the ages (a_0 and a_1) at which the onset rate and progression change. The confidence range for the age at which the onset starts is from 17 to 20 years, and the onset changes between ages 32 and 36. Outside both confidence ranges the fit deteriorates rapidly. Note that in Table 18 in Boyes *et al.* (1982) already a clear difference was shown between estimated incidence rates of dysplasia before and after age 35.

Table IV.2 Maximum likelihood estimates (MLE) and confidence ranges for the parameters of model C with age-dependent progression of pre-invasive lesions.

Parameter	MLE	Range
Incidence of pre-invasive lesions (per 10^3 woman years):		
incidence rate, age < 34 (r_0)	2.11	1.75 - 2.83
incidence rate, age > 34 (r_1)	1.06	0.80 - 1.38
relative risk of participants (rr)	0.74	0.62 - 0.85
Duration and progression of pre-invasive lesions:		
mean duration m (years)	11.8	9.8 - 14.5
shape of distribution (b)	1.58	0.92 - 2.12
progression, age < 34 (p_1)	0.16	0.08 - 0.24
progression, age > 34 (p_2)	0.60	0.42 - 0.88
Pap smear:		
sensitivity (s_p)	0.80	0.76 - 0.85

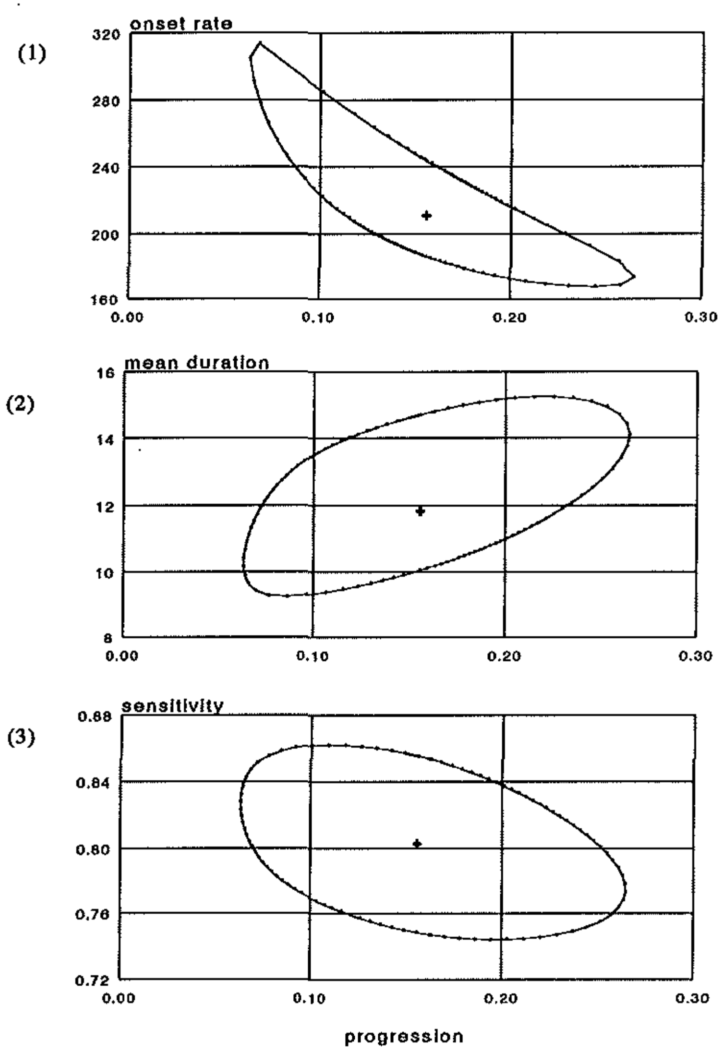

Figure IVb Two-dimensional confidence regions for parameters of pre-invasive cervical cancer, model C. "Progression" is the proportion of progressive lesions among new cases developing before age 34. The points + mark the best-fitting combinations of two parameters.
(1) progression and onset rate of pre-invasive lesions before age 34;
(2) progression and mean duration of pre-invasive lesions;
(3) progression and sensitivity of Pap smear in pre-invasive stages.

Only slightly wider ranges are found when two-dimensional confidence regions are considered, see Figure IV.b. For example, even when the sensitivity would be known to be 77%, then the upper limit for progression is still only 27%. From the figure it can be seen that variation in one parameter may be compensated by changing other parameters as well. For example, a high proportion of regression is possible when the onset rates are high, the mean duration short, and the sensitivity high.

Discussion

In the original analysis of the British Columbia cohort study, it was concluded that regression is part of the natural history of dysplasia and carcinoma *in situ*, and that regression is more likely to occur at younger ages (Boyes *et al.*, 1982). Our analysis combines comprehensive data from this screening project with a modelling approach in which the essential aspects of the natural history of cervical cancer and screening are incorporated. The results show that it is possible to obtain an adequate fit of the screening data set by assuming that regression of pre-invasive lesions is age-dependent, with a higher proportion of regressive lesions developing before age 34. It is estimated that at least three quarters of these lesions in young women and about one-third in middle-aged women will regress spontaneously. We also analyzed a model which 100% progression is assumed in middle-aged women. This will give a statistically satisfactory fit of the cohort study data. However, the fit improves significantly by assuming that regression also occurs in lesions that develop after age 34. Assuming equal regression probabilities at all ages, or no regression at all, will give rise to large discrepancies between the model and the data from the cohort study.

Using Occam's razor in formulating the model has a clear advantage, since it allows for testing of hypotheses in a statistical framework. As a consequence, the resulting model inevitably contains some biologically implausible aspects. For example, age-related changes in the incidence and natural history of pre-invasive lesions will occur gradually, and not briskly at one specific age (34). Thus, the model should be regarded as a filtered version of a more smoothly operating process.

Both the sensitivity and the duration of pre-invasive lesions are assumed to be independent from age. Possibly, there could be more fast growing lesions at increasing age, but testing assumptions about age-dependent duration is hampered by the absence of screening data from age 54 onwards. It is difficult to predict the consequences of such assumptions for the estimated proportion regression, because of the confounding effects of other parameters that will also take different values (as could be seen in Figure IV.b).

The average duration of regressive and progressive pre-invasive lesions is assumed to be equal. A shorter duration for regressive lesions would probably result in an even higher proportion of (new) regressive lesions.

In model (C) with age-dependent regression, the confidence interval for the concentration parameter governing the variability in the duration of pre-invasive lesions includes the value $b = 1.0$ which represents an exponential distribution, see Table IV.2. This means that the current model is just not significantly superior ($0.05 < p < 0.1$) to a further simplified model with exponential dwelling time in the pre-invasive stage. We also considered a log-normal distribution of the duration of the pre-invasive state as an alternative for the Weibull distribution. The resulting parameter estimates are almost exactly the same as those for the Weibull distribution. Only the mean and variation of the duration of the pre-invasive state both give higher values, but it appears that these differences are necessary to have about equal probabilities for durations between 0 and 10 years. The goodness of fit does not improve for model C, and appears to be considerably worse for the models with constant regression or no regression (A and B).

For the preclinical invasive stage we assumed a fixed duration of 4 years and a sensitivity of the Pap smear of 90% in order to arrive at the observed ratio between detection rate and clinical incidence, see Table IVA.4. Other assumptions about preclinical invasive cancers may as well fit the data. For example, a different but still "simplified" assumption is that the duration of pre-invasive and preclinical invasive stages are 100% correlated. This means that lesions with a short dwelling time in the pre-invasive state will also have a relatively short dwelling time in the preclinical invasive stage. It appears that with this assumption, model C results in a equally good fit of the observed data from the cohort study. The values for most parameters are not very different from those listed in Table IV.1. Only for the concentration parameter (b) a different value (2.0) is estimated, but this means that the variability in the duration of the total pre-clinical period is about the same for both assumptions.

There is general evidence that risk and participation to cervical cancer screening are associated (Koopmanschap *et al.*, 1990b). The decision to include the relative risk parameter in the model was supported by the clearly higher clinical incidence rates in the unscreened parts of the cohorts in comparison with the clinical incidence in the total population in the "unscreened" situation in 1955-57. It was noted by Boyes *et al.* (1982) that the accuracy of the clinical incidence rates in the cohorts may suffer from problems in determining the actual size of the unscreened "at-risk" population. However, if these data are not used in the estimation procedure, and the relative risk is given a fixed value assuming either no difference in risk between participants and unscreened women or a relative risk of 0.8 for participants, the resulting parameter estimates are still within the confidence ranges for the full model C as presented in Table IV.2.

Since the period covered by the British Columbia cohort study, there have been clear developments in diagnostic techniques (colposcopy) and follow-up guidelines of early cytological abnormalities. In the early seventies, colposcopy was introduced in British Columbia, hampering continued model-based evaluation of the two study cohorts (Anderson *et al.*, 1988). Given the tendency towards treatment of very early abnormalities, and the impossibility to discern regressive from progressive lesions, it seems probable that the proportion of regressive lesions among those treated after detection by screening will become larger as a result of these developments.

Estimates for the proportion regression based on follow-up of untreated cases of carcinoma in situ show great variations, see Brinton (1986) for an overview. For carcinoma in situ lesions, Kottmeier (1961) reported 71% progression to invasive cancer after 12 years of follow-up. In contrast to this figure is the 36% regression after 5 years of follow-up, as reported by Kinlen and Spriggs (1976). They also found that regression was confined to women aged less than 40 at the time of the initial smear. These follow-up periods are short if compared with the duration of preinvasive stages. The principal value of these studies is thus the support for the existence of spontaneous regression.

Our estimate that a considerable proportion of pre-invasive lesions will progress to invasive stage does not prove a causal relation between dysplasia and carcinoma *in situ* and invasive cervical cancer. Evidence for such a relation is given by the results of a combined analysis of data from major screening programmes (IARC Working Group, 1986), indicating a strong reduction in risk of invasive cervical cancer in the first 5-10 years after one or more negative Pap smears. We have analyzed the IARC data with model C, using the quantification given in Table IV.1. Despite the apparent difference between the long average duration of 11.8 years and the relatively short duration of the protective effect reported by the IARC study, we found that the model gives a good fit of this reduction in risk (see Chapter V).

Among the models for evaluation of cervical cancer screening (see Prorok, 1986, for an overview), a number of other "simplified" models for analysis of screening data have been published. Coppleson and Brown (1975) use data on age specific clinical incidence and detection rates of a first smear. This is a much more limited data set than we used, and their model shows some differences with our model. However, they also found that the possibility of regression should be included in the model in order to explain the observed data.

Albert (1981) tried to fit annual data concerning number of cases with CIS, pre-clinical invasive, and clinical invasive cancer in British Columbia. No distinction is made between first and subsequent smears, and false negatives are neglected. In our opinion, too many important aspects (age-dependencies, difference between first and subsequent smears) are neglected in this model, and it is not suited for testing assumptions about regression of CIS.

Brookmeyer and Day (1987) proposed an extension of the model that Day and Walter developed for breast cancer screening, which is similar to our extension. They analyzed data from a case-control study addressing the question of the relative risk of invasive cancer for women who had a negative smear from. The data come from one of the screening programmes involved in the IARC study (IARC working group, 1986). Detection rates at successive smears are not taken into account, and therefore the proportion of regression can not be estimated from these data. The sensitivity of the Pap smear and the mean duration of the pre-clinical stage are estimated and have a large confidence region that includes our estimates.

Gustafsson and Adami (1989) used a model that is similar to ours, but includes mortality as an additional final stage. Swedish population based incidence (CIS and invasive cancer) and mortality rates are used to obtain estimates for regression and duration of the preclinical stages. The estimated mean duration of the pre-invasive stage is 13.3 years, which compares well with our estimate. Further similarities are found for the variability of the duration (40% of new lesions will become invasive within 10 years, compared to 47% in our model) and the mean duration (3.9 years) of pre-clinical invasive lesions. However, the proportion progressive lesions is estimated to be lower (12%) than in our model, and is found to be independent from age, resulting in a marked difference with our model at higher ages. This low proportion might be due to the fact that in analyzing of the Swedish data, no distinction could be made between results of first and subsequent smears, see van Oortmarssen and Habbema (1990) and Adami and Gustafsson (1990).

Other models for cervical cancer screening are comprehensive rather than simplified, and try to give a realistic description of the processes involved. Such models are less useful for estimation of parameters or testing of hypotheses. Typically, these models aim at evaluation of different screening policies (Knox, 1973; Eddy 1981).

We conclude that the present analysis gives evidence for the existence of a considerable proportion of regression, especially at young ages. The implications of this finding for cervical screening policies can best be considered in a cost-effectiveness framework. Such an analysis, based on this and other model-based analyses of screening data has been carried out for the present situation regarding the epidemiology and early detection and treatment possibilities for cervical cancer. The medical findings are reported in Van Ballegooijen *et al.* (1990), and the economic aspects in Koopmanschap *et al.* (1990a). The results point out that frequent screening at young ages gives rise to an unfavourable balance between favourable and adverse effects. It is also inefficient when comparisons of the cost-effectiveness ratio are made with screening at higher ages.

Appendix IVA.

Table IVA-1. (see page 99)
Overview of the data from the British Columbia cohort study (Boyes *et al.*, 1982), and the fit between expected numbers for models A,B,C and observed numbers: (a) Clinical incidence of invasive cervical cancer 1955-7; (b) Clinical incidence in the unscreened part of the cohorts; (c) First smear: pre-invasive cancers; (d) Second smear by age at midpoint; (e) Third and subsequent smears by age at mid-point; (f) Second smear by interval since first screening; (g) Third and subsequent smears by interval since preceding smear.
* significant difference, $p<0.05$
** significant difference, $p<0.01$

Cases		Observed data		Expected number of cases		
	Age	Rate x 10^5	Cases	Model A	Model B	Model C
(a) Clinical	20-29	3.2	9	8.9	7.9	9.8
Incidence	30-34	10.1	15	20.9	21.1	23.2
	35-39	23.6	36	39.4	40.9	35.5
	40-44	29.8	40	51.2	52.6	43.7
	45-49	45.7	49	51.2	50.1	46.8
	50-54	53.5	43	43.2	38.7	42.0
	55-59	53.6	37	38.9	30.8	40.0
	60-64	55.5	35	35.0	24.6 *	38.4
(b) Unscreened:	25-29	6.8	15	12.3	11.3	14.7
Clinical	30-34	22.7	27	20.8	22.1	24.4
incidence	35-39	53.2	15	17.7	21.3	18.8
	40-44	42.3	85	79.4	82.4	68.5 *
	45-49	56.5	55	59.5	61.7	58.0
	50-54	141.6	39	32.0	32.7	36.3
(c) First smear	20-24	700	7	4.8	5.3	4.7
	25-29	880	106	96.5	104.3	110.5
	30-34	1190	402	367.8	385.4	406.4
	35-39	1150	215	209.9	218.2	220.1
	40-44	1030	150	155.1	158.1	148.3
	45-49	930	223	232.7	229.6	212.9
	50-54	780	81	91.5	88.1	85.6
(d) Second smear	25-29	182	34	32.5	32.2	33.4
by age	30-34	216	24	118.1	114.2	112.7
	35-39	126	25	40.4 **	39.5	33.5
	40-44	125	31	37.1	35.9	31.1
	45-49	122	53	62.4	59.1	52.9
	50-54	109	12	16.8	15.5	14.2
(e) Third + smear	25-29	157	9	7.7	7.5	8.9
by age	30-34	165	96	91.0	86.5	96.3
	35-39	105	63	61.9	64.2	64.2
	40-44	140	16	9.2 *	9.8	10.2
	45-49	101	59	46.7	49.0	52.2
	50-54	70	30	35.6	36.7	40.1
(f) Interval	≤ 11	409	23	30.0	28.2	23.0
second smear	12-35	224	94	107.0	102.4	93.0
(coh 2)	≥ 36	136	66	54.6	55.8	64.2
	≤ 11	270	9	16.2	14.6	10.5
(coh 1)	12-35	150	47	65.8	60.5	49.5
	≥ 36	90	40	34.2	35.4	38.2
(g) Third + smear	≤ 11	248	45	44.7	42.3	38.8
(coh 2)	12-35	119	98	101.3	99.8	107.0
	≥ 36	105	25	14.6 *	16.1 *	23.6
(coh 1)	≤ 11	160	25	27.7 *	26.6	23.6
	12-35	88	66	58.3	60.7	65.7
	≥ 36	63	14	5.5	8.3	13.2

Table IVA-2. Goodness of fit of the three models A, B, C

	Model		
	A	B	C
Goodness of fit			
Chi-Square	73.3	56.0	24.4
D.F.	33	32	31
p-value	0.00007	0.005	0.8
Deviance	69.4	55.2	24.3
p-value	0.0002	0.007	0.8

Table IVA-3. Comparison between models A, B and C.

Models	2 Log Likelihood Ratio (Chi-square)	P-value
A-B	14.2 (1 d.f.)	< 0.0001
B-C	30.9 (1 d.f.)	< 0.0001
A-C	45.1 (2 d.f.)	< 0.0001

Table IVA-4. Comparison of screen-detection rates at first smear, and clinical incidence of invasive cancers.

	Clinical incidence		Screen-detected invasive cancers		
Age	Cases	Rate ($x10^5$) (INC)	Cases	Rate ($x10^5$) (PCI)	Ratio PCI:INC
20-24	1	0.8	0	0	n.a.
25-29	8	5.6	2	17	3.0
30-34	15	9.9	23	68	6.9
35-39	36	22.7	18	97	4.3
40-44	40	27.4	16	110	4.0
45-49	49	40.2	35	146	3.6
50-54	43	45.5	11	105	2.3

Appendix IVB

THE FORMULAE OF THE MODEL.

In this appendix, we will give the formulae that have been used in calculating the expected incidence rates and detection rates as shown in Table IVA.1, on basis of the 10 parameters of the model: a_0, a_1, r_1, r_2, rr, m, b, p_1, p_2, s_P (see Table IV.I).

Functions *f(z)* and *F(z)* are the probability density and distribution, respectively, of the duration *z* of the pre-invasive stage. The Weibull distribution for *F(z)* is characterized by two parameters, scale *c* and shape *b*:

$$F(z) = 1 - e^{-(z/c)^b} \tag{1}$$

The scale parameter *c* can be obtained from the mean *m*, since $m = c\,\Gamma(1+1/b)$, where $\Gamma()$ is the Gamma function.
The probability density and distribution functions for the duration $y=z+q$ of the total preclinical state is denoted by *g(y)* and *G(y)*, respectively, where *q* is the duration of the preclinical invasive stage. The two variants for the relation between *F(z)* and *G(y)* are:

1. Fixed duration *q* of the preclinical invasive stage:

$G(y) = F(y\text{-}q)$, and $g(y) = f(y\text{-}q)$ for $y>q$, $G(y) = 0$ and $g(y) = 0$ otherwise.

2. 100% correlation between *z* and *q*. Here $v = m/(m+q)$ is the average proportion of the preclinical duration in the pre-invasive stage:
$G(y) = F(v.y)$, and $g(y) = v.f(v.y)$ for all $y>0$.

The clinical incidence *I(a)* at age *a* is derived from the onset rate *R(a-y)*, the proportion of progressive lesions *p(a-y)* and the distribution *g(y)* of the total duration $y=z+q$ of the two pre-clinical stages:

$$I(a) = \int_0^{a-a_0} g(y)p(a-y)R(a-y)dy \tag{2}$$

The clinical incidence applies to the total female population at risk, which by definition has a relative risk equal to 1.0. For known values of the relative risk *rr* of the screened population, the relative risk of the unscreened population *ru(a)* is age-dependent via the fraction screened $\pi(a)$:

$$ru(a) = \frac{1-\pi(a).rr}{1-\pi(a)} \tag{3}$$

The incidence in the unscreened part of the cohorts is *ru(a) . I(a)* and can be derived from expressions (2) and (3).

For convenience, we introduce *N(a)*, the rate at which cases leave the PRE-INVASIVE stage:

$$N(a) = \int_0^{a-a_0} f(z) R(a-z) dz \tag{4}$$

Now the detection rate of pre-invasive lesions at a first screening at age *a* is:

$$P_P(a) = rr\ s_P \int_0^a \big(R(x) - N(x)\big) dx \tag{5}$$

And the detection rate of invasive pre-clinical lesions:

$$P_I(a) = rr\ s_I \int_0^a \big(N(x) - I(x)\big) dx \tag{6}$$

The detection rate of pre-clinical lesions for a second smear at age u_2 depends on the age u_1 at which the first smear was made.

$$S(u_1,u_2) = \int_{a_0}^{u_2} rr R(a) \Big[s_P \big\{ 1 - F(u_2 - a) \big\} + s_I p(a) \big\{ F(u_2 - a) - G(u_2 - a) \big\} \Big] (1 - s_P)^{a<u_1} da \tag{7}$$

The notation $(1-s_P)^{a<u_1}$ is used to indicate that the false negative rate at the first screening should be taken into account only in cases where the onset occurred before age u_1.

For the detection rates of second and later smears, a further simplifying assumption was made to reduce numerical complexity: the false negative rate at the subsequent smear(s) is assumed to be $1.0{-}s_P$ for all screen-detectable lesions. Thus, the expression for the false negative rate is an approximation, neglecting the lower false negative rate $1.0{-}s_I$ in cases who are in stage PRECLINICAL INVASIVE already at age u_2. This will result in a slight exaggeration of the detection rate, since lesions that are in stage PRE-CLINICAL INVASIVE CERVICAL CANCER at preceding smears would have a false negative rate of u_2.

The detection rate at a third screen at age u_3 depends on the ages u_1 and u_2 of the preceding smears:

$$
\begin{aligned}
T(u_1,u_2,u_3) = & \int_{a_0}^{u_1} rr\ R(a)(1-s_P)^2 \left[s_P\{1-F(u_3-a)\} + s_I p(a)\{F(u_3-a) - G(u_3-a)\}\right] da \\
& + \int_{u_1}^{u_2} rr\ R(a)\ (1-s_P) \left[s_P\{1-F(u_3-a)\} + s_I p(a)\{F(u_3-a) - G(u_3-a)\}\right] da \\
& + \int_{u_2}^{u_3} rr\ R(a) \left[s_P\{1-F(u_3-a)\} + s_I p(a)\{F(u_3-a) - G(u_3-a)\}\right] da
\end{aligned}
\qquad (8)
$$

The simplifications in *R(a)* and *p(a)* are:

$$
R(a) = \begin{cases} r_1 & a_0 < a < a_1 \\ r_2 & a > a_1 \end{cases}
$$

$$
p(a) = \begin{cases} p_1 & a_0 < a < a_1 \\ p_2 & a > a_1 \end{cases}
$$

Expressions (2), (5), (7) and (8) can now be simplified considerably. The resulting expressions are used to calculate expected rates for a given set of parameter values. The clinical incidence *I(a)* and the detection rate of a first smear are calculated for 1-year age-groups and then aggregated to the classes used in the testing procedure. The detection rates $S(u_1,u_2)$ of a second smear are calculated for a matrix of "ages at mid-point of the interval" and intervals. The same method is used for detection rates $T(u_1,u_2,u_3)$ of third and subsequent smears, assuming that the interval between the first and second smear is 1, 2, 3 or 5 years with probabilities 0.40, 0.30, 0.15 and 0.15, respectively (based on woman-years in the published tables for the second smear). These rates are also aggregated, after dividing the rates by the length of the interval to obtain rates that have woman-years as denominator. Expected numbers of cases are obtained by multiplying expected rates and observed denominators (woman-years or number screened).

The log-likelihood is based on the assumption that the observed cases are a realisation of a Poisson-distribution with mean = expected number of cases. The likelihood is maximized using a downhill simplex multidimensional optimization routine (Press et al, 1988).
The total number of classes is 45, the number of degrees of freedom for the Pearson Chi-Square test for the goodness of fit of the model is 41 minus the number of free parameters that were varied in deriving the maximum likelihood estimates. The

number of degrees of freedom equals the number of categories minus 4, since in each cohort both expected and observed sum of cases detected with the second smear and with subsequent smears is the same for the two subclassifications (by age and by interval since first smear), as can be seen in Table IVA-I. The deviance, i.e. the likelihood ratio test statistic for comparing a model with the complete model, is also inspected in assessing the goodness of fit.

Comparisons between models are based on the Likelihood-ratio test. Also, 95% confidence regions for 1 and for two parameters are obtained by inverting the Likelihood-ratio test, i.e. by searching for parameter values for which the log likelihood is 3.84 ÷ 2 respectively 5.99 ÷ 2 lower than the log-likelihood of the optimal model. One- and two-dimensional confidence regions (Table IV.II, Figure IV.b) are computed by repeatedly applying the downhill simplex optimization routine in combination with a root-finding algorithm.

V. THE DURATION OF PRE-CLINICAL CERVICAL CANCER AND THE REDUCTION IN INCIDENCE OF INVASIVE CANCER FOLLOWING NEGATIVE PAP-SMEARS.

Introduction

The results of three decades of cervical cancer screening have been analyzed by different methods in order to assess the effectiveness of early detection and treatment, and to make recommendations about the screening intervals. The most convincing direct evidence for the effectiveness of screening is given by the mortality trends in countries with different levels of screening (Läärä *et al.*, 1987). However, such a comparison does not allow conclusions about the optimal screening interval. This is an important question in view of both the large resource implications of cervical cancer screening, and the striking diversity in policy recommendations.

In analyzing the interval between Pap-smears, two methods are being used: a direct epidemiologic exposure-incidence approach and a more indirect model-based approach.

In the direct approach, the potential impact of different screening intervals is derived from the gradual increase of the incidence of invasive cervical cancer in women after one or more negative Pap-smears. This incidence can be estimated directly by linking screening register data and cancer register data, an approach which has been used by an IARC working group in a combined analysis of results from large screening programs in 8 countries in Europe and North America (Hakama *et al.*, 1986, IARC, 1986). The IARC group reported the results as "relative protection", i.e. the ratio of the incidence in unscreened women to the incidence following negative smears. We prefer the term "risk reduction" to avoid the suggestion of active protection given by Pap-smears. In the first years after a negative Pap smear, the relative risk is less than 25 percent of the level in unscreened women, and within 6-10 years it increases to approximately 50 percent of the unscreened level (Hakama *et al.*, 1986).

The second approach uses a "deep" model of the pre-clinical course of the disease, which is fitted to screening results, and subsequently used to predict the favourable and adverse effects of screening policies. Examples include Coppleson & Brown (1975) and Gustafsson & Adami (1989,1990,1992). We have developed a model to analyze the epidemiological data and the detailed screening results from British Columbia, see Chapter IV. One would expect that the duration of low risk in the 'risk reduction' approach would be approximately equal to the duration of the pre-clinical stages in the model-based approach. However, model-based estimates are consistently higher, with a mean total duration of dysplasia, carcinoma in situ (CIS), and pre-clinical invasive stages of 15 to 20 years (Coppleson and Brown, 1975, Gustafsson and Adami, 1990, and Chapter IV) and a median of about 15 years.

In the IARC study both clinically diagnosed cancers and screen-detected invasive cancers have been used in determining the risk reduction, see page 17 of the report (Hakama *et al.*, 1986). We will investigate whether this inclusion of screen-detected cancers can give an explanation for the seemingly contradictory outcomes of the two approaches. The implications for the choice of the screening frequency will be discussed.

Material and methods

Our epidemiologic model describing the pre-clinical course of cervical cancer and the intervention by screening, which was used to analyze comprehensive data from the British Columbia cohort study of cervical cancer screening, will be used to analyze the incidence of invasive cancer following negative smears as calculated in the IARC study.

model.

The British Columbia cohort study (Boyes *et al.*, 1982) is based on cervical screening data collected between 1949 and 1969, in two cohorts born between 1914-18 and between 1929-33, respectively. The parameters of our cervical cancer screening model were estimated from the cohort study data, which include age-specific detection rates at first screening and at subsequent screenings by interval since preceding smear, the clinical incidence in the population before screening started, and the clinical incidence in unscreened women. The resulting model quantification shows a good fit of the data, see Chapter IV.

The model consists of six stages, see Figure Va. The model neglects death from cervical cancer and from other causes and hysterectomy for other reasons than cervical cancer. These stages can be excluded, since comparison between observed data and outcomes of the model are based on rates, and it is assumed that the risk of death from other causes or hysterectomy for other reasons than cervical cancer is independent from the risk of cervical cancer.

From the initial stage NO CERVICAL CANCER, transitions to the stage PRE-INVASIVE CERVICAL CANCER (which includes Dysplasia and Carcinoma in Situ) occur at two different rates, for younger and for older women, respectively. The duration of this stage is governed by a probability distribution function which is characterized by the mean duration (estimated value: 12 years) and the variability of this duration which is considerable: 20% of new progressive lesions will enter the stage PRE-CLINICAL INVASIVE CANCER within 5 years. The estimated probability of spontaneous regression is 84% for younger women and 40% for older women.

In the absence of screening, all invasive cancers will eventually be diagnosed clinically (stage CLINICAL INVASIVE CERVICAL CANCER). The duration of PRE-CLINICAL INVASIVE cancers is assumed to be constant, and estimated to be 4 years. Thus, the total average duration of (progressive) screen-detectable stages is between 14 and 20 years. The probability of detecting a pre-clinical lesion by a Pap-smear is estimated to be 80% in PRE-INVASIVE CERVICAL CANCER and 90% in PRE-CLINICAL INVASIVE CANCER. Screening participants constitute a group which has below-average risk of developing cervical cancer; the estimated relative risk is 74%. Confidence regions of the parameter estimates and other details of the analysis of the British Columbia data and the resulting model are given in Chapter IV.

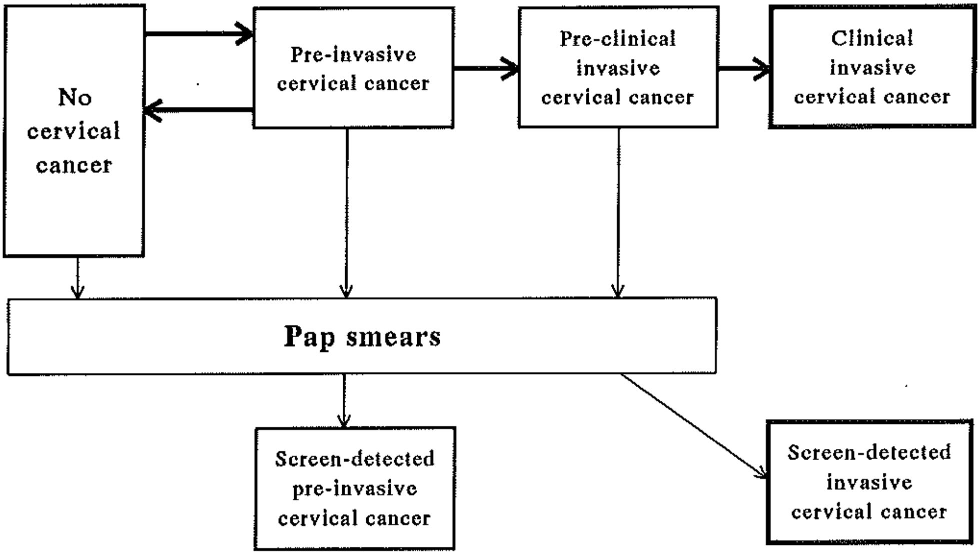
No cervical cancer
Pre-invasive cervical cancer
Pre-clinical invasive cervical cancer
Clinical invasive cervical cancer
Pap smears
Screen-detected pre-invasive cervical cancer
Screen-detected invasive cervical cancer

Figure Va. Schematic representation of the cervical cancer screening model

The model is used to predict the build-up of incidence of invasive cancers following negative smears which can be compared with the estimates of the IARC study for the relative protection against invasive cervical cancer following one negative smear and two or more negative smears, respectively (see Tables III and IV in IARC, 1986). This relative protection was calculated by dividing the incidence in unscreened populations (taken from Cancer Incidence in Five Continents, volumes II and III) by the observed incidence of invasive cervical cancer in screened women in the screening programs involved in the IARC study.

Comparing the model with IARC findings

The model has been used to predict the initial fall and subsequent rise in incidence of invasive cervical cancer following a first or second negative smear. Women who had negative Pap smears are likely to have further preventive smears at which invasive cancers may be detected in addition to the clinically diagnosed (symptomatic) cancers. In the studies that are based on complete cohort data, the total incidence of invasive cervical cancer is a function of age and time since last negative smear, number of negative smears, and screening intensity after the last negative smear. The numerator of the incidence is composed of the clinically diagnosed cancers and the product of the screening intensity and the screen-detection rate of invasive cancers (see Appendix).

The main results from the IARC study concern the relative risk following two or more negative Pap smears in women aged 35-64. The interval between two smears is approximately 3 years in most of the programs that participated in the IARC study, with a range from 2-3 years to 5 years (Hakama *et al.*, 1986). We assumed an interval of 3 years between the last 2 negative smears. The contribution of screen-detected invasive cancer to the total incidence depends on the screening intensity for which we assumed that each year 33 percent of the women will have a Pap smear. Reports of the participating programs indicate that the average age of the women in age-group 35-64 is approximately 45 years (Hakama *et al.*, 1986). Therefore, model predictions for age 45 will be compared with IARC outcomes, which are predominantly based on complete cohort data (114 of the 162 cases of invasive cancer, or 70%). The remaining 48 invasive cases are from case-control studies. In principle, these case-control data would require a separate treatment in the model because of the different consequences of pooling screen-detected and clinical cases, and the absence of high-risk non-participants in these studies.

IARC results for the relative risk following a single negative smear have only been presented for women below age 35 (IARC, 1986), and are fully based on cohort data. Most women concerned are in the 30-34 age-group. We take screen-detected invasive cancers at a second smear into account, again assuming that each year 33 percent of the women will have a next Pap-smear.

Results

In *Figure Vb(1),* the risk reduction following two negative smears as predicted by the model is compared with the observed results of the IARC study (women aged 35-64). Confidence ranges for the IARC outcomes are given. The predicted relative risk of clinically diagnosed cervical cancer is much lower than the observed IARC level. During the first 4 years the predicted risk is very low because of the assumed fixed 4-year duration of preclinical invasive stage and the 10% false negative rate in this stage at each screening. At 4 years, an abrupt (small) increase in risk occurs because of the higher false-negative rate (20%) of pre-invasive lesions, and from this moment also cervical cancers which started developing after the last smear can be diagnosed. Note that part of the risk reduction predicted by the model results from the difference in risk between participants and non-participants, which would give a relative risk of 74% even if Pap-smears had no effect at all.

Much better agreement between model predictions and IARC outcomes is obtained when screen-detected invasive cancers are included in the prediction. The contribution of screen-detected invasive cancers in year *y* is equal to the prevalence of preclinical invasive cancers (which is roughly 4 x the average clinical incidence in years *y* to *y+4* because of the 4-year duration of the preclinical invasive stage), multiplied by the screening intensity (0.33) and the sensitivity (0.9). Note that this rather high screening intensity will lead to a rapid decline in the number of women-years at risk after the first few years. In these first years, the predictions are slightly too low, which might be due to the simplified model assumption of the fixed duration of the pre-clinical invasive stage. Comparison for other ages within the 35-64 range shows that the predicted relative risk decreases gradually with age. We conclude therefore that the predictions for age 45 as shown in figure Vb(2) probably give a good account of the build-up of incidence following negative smears in the age-group 35-64.

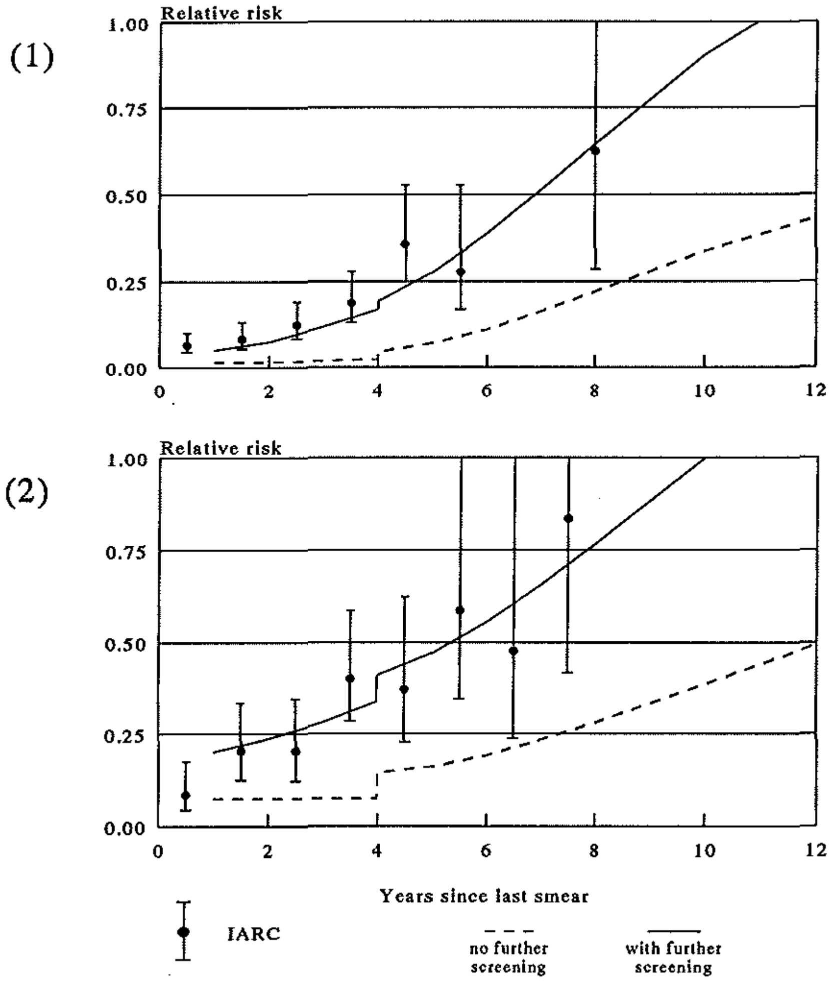

Figure Vb. Comparison of estimated relative risks (mean and 95% confidence interval) of cervical cancer according to the IARC study (IARC, 1986), and model predictions based on the best fitting estimates of the British Columbia data: (1) women aged 35-64 who had two or more negative smear; (2) women under 35 years who had one negative smear. The IARC estimates in (1) and (2) are adapted from Table V.3 and V.4 in (IARC, 1986).

no further screening:	only clinically diagnosed cancers, no further screening.
with further screening:	both clinically diagnosed cancers and invasive cancers detected at further screening.

In *Figure Vb(2)*, the model predictions and IARC outcomes are compared for women below age 35 who had had one negative smear. Because of the smaller impact of a single negative smear, the relative risks in the first years after the negative smear are higher than in *Figure Vb(1)*. The predicted trend for clinically diagnosed cases starts at a relative risk of 7.4% in the first four years following the negative smear which results from the 10% false negative rate and the 74% relative risk of participants. The model-predicted relative risk of clinically diagnosed cancers is lower than the IARC outcomes, and is below the lower limit of the confidence interval for nearly all IARC outcomes. Inclusion of screen-detected invasive cancers will give a much better agreement. The model predictions are shown for age 35. For younger ages higher relative risks are predicted, because of the larger proportion of fast growing lesions at these ages. The predictions for age 30 are well within the confidence ranges of the IARC outcomes, however.

Discussion

The results of the model can be used to draw conclusions about the risk reduction after negative screening tests, and about the duration of (progressive) preclinical stages of the disease.

A first conclusion is that the estimated relative risks of the IARC study should be interpreted with caution because of the inclusion of screen-detected invasive cancers. These cases cause an early increase in the risk of invasive cancer, as can be seen from the predicted curves "no further screening" and "with further screening" in figure 2. The difference between the curves depends only on two assumptions: the duration of pre-clinical invasive cancer, and the frequency of screening.

The 4-year duration of pre-clinical invasive cancer is derived directly from comparison of detection rates at first screening and clinical incidence in unscreened women in British Columbia, see Chapter IV and Boyes *et al.* (1982). Similar durations have been obtained in Sweden (Gustafsson and Adami, 1992) and from a large pilot study of cervical cancer screening in The Netherlands.

From equations (1) and (4) in Appendix VA it can be seen that the difference between "clinical only" and "all invasive" is proportional to the screening intensity. Given the mix of the different screening intervals obtained by pooling the data from the separate programmes in the IARC study, an average frequency of 33% per year between 2 and 6 years after the last negative smears is reasonable.

A second conclusion is that accounting for the impact of further screening can resolve the apparent discrepancy between the results of the IARC study, which might suggest a median duration of 5-8 years of pre-clinical cancer, and estimated median durations of 15 years and more derived with statistical models. This conclusion is based on our statistical model, with parameters which had been estimated directly from comprehensive screening data from British Columbia.

The only parameter that had to be added to the British Columbia model in analyzing the IARC data is the screening frequency. We tested the implications of alternative assumptions about the interval between smears. Other intervals, for example 2 or 4 years in stead of 3 years, will still result in a good correspondence with IARC findings and model predictions concerning relative risk after two negative smears, provided that similar assumptions are made for the duration between the two smears and the duration between the second and a possible subsequent third smear. A shorter interval between the two negative smears will give more reduction in the clinical incidence of invasive cancers, but this is compensated by the associated higher screening frequency following the second smear which will increase the number of screen-detected invasive cancers. Similarly, a longer first interval can be compensated by a lower screening frequency following the second smear. The predicted relative risk after one negative smear is more sensitive to changes in assumptions about a possible second smear. Assuming that each year only 25 percent of the women (who had a first negative smear) are screened would lead to a lower value of the predicted relative risk, but most values remain within the confidence limits of the IARC data. If 50 percent of the women would have a smear each year, then the predicted relative risk is clearly too high, especially in the first 5 years following the negative smear.

A restriction with respect to the comparison for two or more negative Pap smears (Figure Vb(1)) is that the way in which the model results are calculated applies to the 70% of the cases that stem from cohort-based incidence data. The remaining 30% of cases stem from two case-control studies in which the relative risk is calculated as the average risk for screen-detected and clinically diagnosed cases, in stead of taking the sum. This would lead to a lower model prediction for this group of cases, which would however be partly compensated by deleting the attenders risk of 74% which is not applicable since persons that have not been screened are not included in the case-control studies.

We also tested the robustness of our second conclusion, that the apparent discrepancy between the results of the IARC study and the estimates of statistical models can be resolved by accounting for the impact of further smears, by investigating alternative quantifications of our statistical model that still give a good fit of the British Columbia data.

The assumption of a fixed 4 year duration of pre-clinical invasive cancers is a simplification. In reality, this duration will show variation, and might be correlated with the duration of the pre-invasive stage. We investigated the extreme opposite assumption that the duration is fully correlated with the duration of this preceding stage. In this case, the average duration of the invasive stage is 5.7 years. The only parameter which is seriously affected by this alternative assumption is the variability of the duration of the pre-invasive stage: only 10% of new progressive lesions will become invasive within 5 years. This alternative model gives an equally good fit of the British Columbia data see Chapter IV. The alternative model predicts relative risks of clinically diagnosed cervical cancer following one or two negative smears that are similar to those predicted by the basic model, and are again clearly below the observed outcomes of the IARC study. In the case of one preceding negative smear, inclusion of screen-detected cancers will result in a good agreement between the model-predicted and the observed risk reduction. For two negative smears, the predicted relative risk shows a larger discrepancy and is 20-30 percent too low if compared to the IARC results. A much better fit can be obtained by increasing the standard deviation of the duration from 5.9 to 7.3 years, which implies more fast-developing lesions. A standard deviation of 7.3 years is still within the confidence limits derived from the British Columbia data.

Another simplifying assumption in the model concerns false-negative test results. Test results in a woman are assumed to be independent, whereas in practice it is likely that after a false-negative smear the sensitivity of a subsequent test is relatively low. After two negative smears (Figure Vb(1)), this would imply a higher starting level of the relative risk. The change in the contribution of screen-detected invasive cancers will depend on the precise assumptions made about the association between test results.

The self-selection relative risk r=0.74 is a model estimate based on the British Columbia data. It is mainly based on ratio of incidence in unscreened women to incidence in historical controls, and as such, can be questioned. A further question is whether it would apply to other regions and countries as well.

The model quantifications for duration and progression closely resemble estimates from a model which was based on data from the Swedish cancer registry on carcinoma in situ (CIS) and invasive cancers (Gustafsson & Adami, 1989). This model only considers CIS, but in the Swedish classification practice this probably includes most cases of severe dysplasia as well. In our model the pre-invasive stage represents both CIS and dysplasia, but dysplasia only plays a minor role in British Columbia in the period considered. In the Swedish model, the estimated average duration of progressive CIS is 13-14 years with a confidence interval of 4 years, and approximately 40 percent of new progressive pre-invasive lesions will become invasive within 10 years, which is nearly equal to the value (39 percent) for our alternative model. The part of the Swedish model which describes progressive lesions is very similar to our model. Differences occur in the proportion and the duration of regressive lesions, but they do not influence the estimated reduction in incidence of invasive cancer.

The assumptions of our model differ markedly from the assumptions made in the cervical cancer model of Eddy that has been claimed to be compatible with the IARC results (Eddy, 1987). This model has not been fitted to actual screening data, but was composed on basis of literature data and has been used for making recommendations about screening policies (Eddy, 1981; Eddy, 1990). Apparently, no provision is made for pre-clinical invasive cancers that are likely to be detected at screening tests. The average duration of the pre-clinical stage is 8 years, which is only half of the estimated value derived directly from detailed data analyses by our model and the model of Gustafsson.

In the introduction we described two approaches to making inferences about the natural history of pre-clinical cervical cancer and the difference in their outcomes: a direct exposure-incidence approach, and a model based approach. Our analysis shows that the seemingly conflicting results of the two approaches can be explained by the observation that women who have had negative smears are likely to have subsequent smears at which invasive cancers can be detected early. Indeed, for women without further smears, the model predicts a relative risk of at most 40% after 10 years following two negative smears, whereas inclusion of screen-detected invasive cases will after 10 years result in a risk which is approximately equal to an unscreened population (see Figure Vb(1)).

The "exposure-incidence" approach has the convenience of giving a direct estimation of risk of invasive cancer following negative smears from epidemiologic data. Uncertainties about some of the parameters of the model-based approach, notably the proportion of regressive cases among screen-detected pre-invasive lesions, are avoided in this approach. The problem, however, is that the risk estimates may be severely biased, depending on the actual frequency of subsequent Pap-smears.

Morrison (1982) already discussed this potential bias in the outcomes of case-control studies of cancer screening, emphasizing that case definition should be based on disease manifestations that develop only after the lead-time interval. Including all screen-detected invasive cases will give rise to too pessimistic estimates of the efficacy of screening. For example, the IARC group stated that "the interval between screening should be three year or less" (IARC, 1986). This recommendation is based on the consideration that an acceptable interval should at least give a 90 percent reduction in the cumulative incidence of invasive cervical cancer in the age group 35-64. However, a considerable proportion of these invasive cancers is detected by screening, and have a much better prognosis than clinically diagnosed cancers. The outcomes of the "exposure-incidence" approach can therefore not be applied to situations with a different intensity of screening, or for making recommendations about screening intervals.

Exclusion of the screen-detected cases from the analysis is not a very useful alternative, not only because the mode of detection is not known in part of the programmes that participated in the IARC study, but especially because the results would then only apply to situations in which screening is stopped.

The model based approach for epidemiologic evaluation of cervical cancer screening has two advantages. First, it takes more epidemiologic data into account in analyzing results of screening programs. The resulting model quantification will thus be compatible with detection rates of first and subsequent screening tests, incidence of interval cancers, incidence in unscreened women (Chapter IV). Indeed, it is an explanatory model for the findings of a screening program, that can be used to show the additional reduction in risk of death from cervical cancer which is incurred by the favourable prognosis of screen-detected invasive cancers. The second, related, advantage is that the model can be used for making predictions of the effects of screening policies, in which both the prevention of invasive cancer and the early detection of (micro-)invasive cancer contribute to the favourable effects by choosing mortality as end-point. On basis of data regarding the survival of screen-detected cancers, our model predicts that a 90 percent reduction in mortality may be achieved with screening intervals of 5 years, see Chapter VI. A detailed cost-effectiveness analysis based on an extended version of the model presented in this paper, indicates that reducing the screening interval to three years or less gives an unfavourable shift in the balance between beneficial effects and adverse effects when compared with longer screening intervals (Van Ballegooijen *et al.*, 1992), and are also far less efficient in terms of cost-effectiveness (Koopmanschap *et al.*, 1990).

Appendix V.A

CALCULATION OF THE REDUCTION IN INCIDENCE FOLLOWING NEGATIVE SMEARS.

Mathematical descriptions of screening for cancer have been presented before, both for one stage and two-stage models (Louis *et al.*, 1978; Day & Walter, 1984; Brookmeyer & Day, 1987; Alexander, 1989). The following description applies to relative risks that are calculated on the basis of data from a complete cohort, i.e. from incidences based on number of cases and number of woman-years.

Consider women who have had one negative smear. The total reported incidence of invasive cervical cancer $I(a,y)$ for women of age a, at y years following the negative smear, is the sum of the clinical incidence $C(a,y)$ and the detection rate of invasive cancers at a second smear $D(a,y)$ multiplied by the screening intensity k (smears per woman per year) in the period since the negative smear:

$$I(a,y) = C(a,y) + k\, D(a,y) \qquad (1)$$

For women of age a who did not yet have a second smear, the clinical incidence $C(a,y)$ at y years after one negative smear is:

$$C(a,y) = \left(1 - s(y)\right) \int_{0}^{a-y} r\, R_p(t)\, g(a-t)\,dt + \int_{a-y}^{a} r\, R_p(t)\, g(a-t)\,dt \qquad (2)$$

where $g(t)$ is the probability density function of the total preclinical duration, and $R_p(z)$ the onset rate of progressive pre-invasive lesions at age z, and $s(y)$ the test sensitivity of the smear at y years before clinical diagnosis. If $y \leq 4$ years, then the cancer was already invasive and $s(y) = s_I$, if $y > 4$ years the cancer was pre-invasive and $s(y) = s_P$. Parameters $s_P = 0.8$ and $s_I = 0.9$ denote the sensitivity of the Pap smear in detecting pre-invasive and invasive lesions, respectively. A relative risk r is introduced to reflect the lower incidence of cervical cancer in participants because of selective participation in screening. The first part of the equation represents the contribution of false-negatives, the second part represents lesions that developed after the negative smear.

At *y* years after the negative smear, the detection rate of invasive cancers *D(a,y)* at a second smear made for women of age *a* is:

$$D(a,y) = s_I\big(1-s(y)\big)\int_0^{a-y} r\,R_p(t)\,\big(F(a-t)-G(a-t)\big)dt \; + \\ + \; s_I\int_{a-y}^{a} r\,R_p(t)\,\big(F(a-t)-G(a-t)\big)dt \qquad (3)$$

where *F(t)* and *G(t)* are the probability distribution functions of the durations of the pre-invasive and the total pre-clinical stage, respectively.
In a situation of "no further screening", the relative risk is the ratio of the clinical incidence *C(a,y)* following a negative smear to the clinical incidence *B(a)* in unscreened women of the same age *a*:

$$B(a) = \int_0^{a} R_p(t)\;g(a-t)dt \qquad (4)$$

If screen-detected cases are included, the relative risk is the ratio of *I(a,y)* to *B(a)*. Similar equations are used for the case of two preceding negative smears. These equations are used for the general case of women who had at least two negative smears; the effect of the additional negative smears is assumed to be negligible.

VI. PREDICTING MORTALITY FROM CERVICAL CANCER AFTER NEGATIVE SMEAR TEST RESULTS

Introduction

The International Agency for Research on Cancer (IARC) working group on evaluation of cervical cancer screening programmes has analysed data from large screening programmes in Europe and North America (Hakama *et al.*, 1986). Protection against cervical cancer after negative smear test results was measured by the risk of invasive cancer. The relative protection - that is, the ratio of the risks in unscreened and screened women - decreases with increasing time since the last smear. Analysis of this decreasing trend may be useful in deciding about the interval between smears. The IARC group concluded that the intervals between screening should be three years or less (IARC, 1986).

The primary aim of cervical cancer screening is to prevent death from cervical cancer. Using death rather than invasive cancer as the criterion will result in different estimates for the protection after negative smear test results. Women who have had negative smear test results are likely to have further smear tests, and invasive cancers may not only surface clinically but also be detected by these subsequent smears. On average, these invasive cancers detected by screening have a much better prognosis than those that are clinically diagnosed. When comparing women who had one or more negative smear test results with unscreened women, the relative protection against mortality from cervical cancer will therefore be higher than the relative protection against invasive cancer. We investigated the size of the difference between these two relative protection rates, and the implications for recommendations about the screening interval.

Methods

The aim of screening for cervical cancer is to reduce the risk of diagnosing an invasive cancer that would eventually result in death from the disease. The next screening should therefore be scheduled when the risk of lethal invasive cancer is considered to have become too high. We therefore used protection against lethal invasive cancer as an indicator for protection against mortality.

Information on the survival of patients with invasive cancers was not available in the IARC study. We therefore used a model for progression of the disease which links the risks of lethal invasive cancer to the data on risk of invasive cancer obtained in the IARC study. We constructed a model which has four stages: no cervical cancer, preinvasive cervical cancer, pre-clinical invasive cervical cancer, and clinical invasive cervical cancer. We used data from the cohort study of cervical cancer screening in British Columbia (Boyes *et al.*, 1982) to estimate the parameters of this model (Chapter IV). The preinvasive stages dysplasia and carcinoma in situ together had an estimated mean duration of 12.3 (95% confidence interval 9.7 to 14.7) years with a standard deviation of 5.8 (4.6 to 11.1) years. The estimated mean duration of preclinical invasive stages was five years. The sensitivity of the smear test was estimated at 80%. These values fitted the data from British Columbia well, and agreed with similar analyses of more recent screening data from the Netherlands (Habbema *et al.*, 1988) and Sweden (Gustafsson and Adami, 1989).

To be consistent with the IARC study, we included screen detected invasive cancers in the calculations for the model. On the basis of the IARC report we estimated that each year 33% of women who have already had two or more (negative) smear test results will have another smear test (Hakama *et al.*, 1986). We used this estimate in the model, to predict the relative risk of invasive cancer; these predictions agreed with the IARC group finding of 50% reduction in risk after 6-10 years after (at least) two negative smear test results. The fit of the IARC results could be improved further by increasing the standard deviation of the duration of preinvasive lesions to 7.3 years, which is still well within the confidence limits of the British Columbia estimates, see Chapter V. The resulting best fitting model was used to translate the protection against invasive cancer P_I into protection against lethal invasive cancer P_M. The model gives the proportion *(s)* of the invasive cancers that are screen detected. On the basis of estimates of the average lethalities q_s and q_c of screen detected and clinically diagnosed invasive cervical cancers, we used the following formula to calculate the protection against lethal invasive cancer:

$$P_M = \frac{q_c}{(1-s)\, q_c + s\, q_s}\, P_I$$

We assumed a relative five year survival rate of 90% for screen detected invasive cancers, which we derived from combining stage distribution of screen detected invasive cancers and stage specific five year survival data (Van der Graaf, 1987). Assuming a 60% five year survival for clinically diagnosed cervical cancers gives a 1:4 lethality ratio (q_s/q_c=0.25). Some studies have reported survival statistics for clinically diagnosed cancers in the range

65%-70% (Van der Graaf, 1987; Cancer Registry of Norway, 1980; Hakulinen *et al.*, 1983), but these figures are higher because stage IA and IB cancers detected by screening were included. We assumed that the 1:4 lethality ratio after 5 years of follow up persisted in the long term; we also considered a lethality ratio of 1:3.

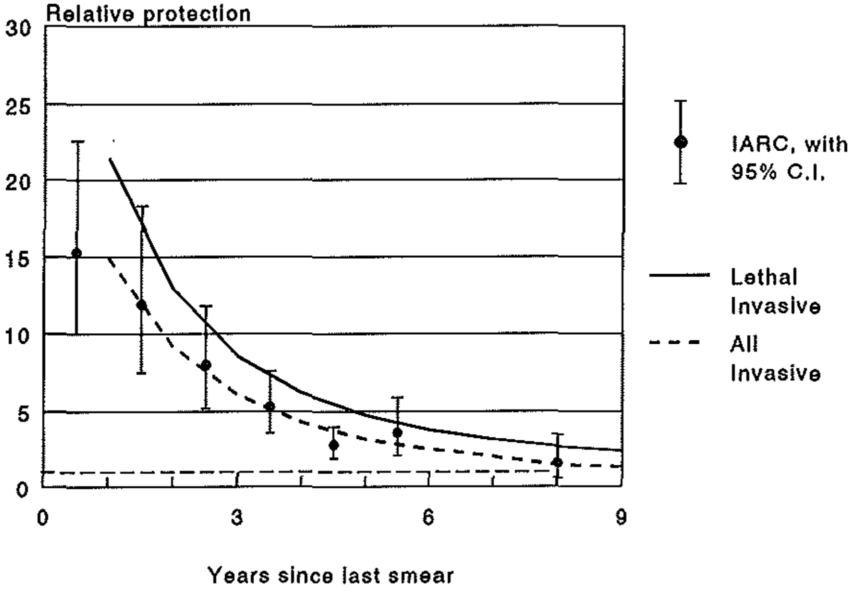

Figure VIa Comparison of relative protection against all invasive cervical cancers as estimated by the IARC group (IARC, 1986) with model predictions of protection against lethal invasive and against all invasive cervical cancer, after two negative smear test results (three years apart) in 45 year old women. It is assumed that each subsequent year 33% of women who have not yet had a third smear will have one. The lethality ratio of screen detected to clinically diagnosed cancers is assumed to be 1:4.

Results

Figure VIa shows the relative protection against invasive cancer and against lethal invasive cancer predicted by our model for a 45 year old woman who has had two negative smear test results at an interval of three years assuming a

1:4 lethality ratio between screen detected and clinically diagnosed cancers. The points give the relative protection against invasive cervical cancer for 35-64 year old women estimated by the IARC study group (IARC, 1986). The dashed curve shows the model fit of the observed results. The model predicted protection against mortality from cervical cancer after negative smear results was about 50% higher than the protection against invasive cancer. If a more conservative lethality ratio of 1:3 was assumed, the relative protection against mortality from cervical cancer was still about 30-40% higher than the IARC estimates would suggest.

In the model, the durations of preclinical invasive lesions have an inherent one to one relation with the durations of preinvasive lesions. This assumption implies that fast growing preinvasive lesions will also develop rapidly when they have become invasive. Screening tends to pick up slow growing lesions and to miss fast growing lesions (length biased sampling). Shortly after two negative smear results some women will already have fast growing lesions that may later be diagnosed clinically, resulting in a small proportion of screen detected cancers and a low prediction of the relative protection against lethal invasive cancers (see formula). This effect is strongest in younger age groups. In the model, most cervical cancers occurring at younger ages are assumed to be fast growing lesions that have a lower probability of being detected by regular screening. Thus, the relative protection against lethal invasive cancer in women under age 30 is predicted to be much lower than for women between ages 35 and 65, for whom predictions are similar to those presented in the figure.

We also considered a model with a fixed four year duration of the preclinical invasive stage. The fit for the two data sets (British Columbia and IARC) was as good as with the former model. The length biased sampling phenomenon did not occur in this model, and therefore it predicted that a higher proportion of invasive cancers will be detected by screening. As a result the relative protection against lethal invasive cancer in the first years since the last negative smear test result was almost 100% higher than the protection against all invasive cancers.

In reality, the development of preclinical invasive cancer will be in between the extreme situations assumed in the two models and the relative protection against lethal invasive cancer will be between 50% and 100% higher than for all invasive cancers.

Screening frequencies

The IARC working group calculated the reduction in cumulative incidence of invasive cervical cancer for different screening frequencies in women aged 35-64. It assumed that all women participate in screening. For example, the figure shows that the incidence of all invasive cancers in the first three years after a second negative smear test result is between 10 and 11 times lower on average than in a situation without screening resulting in an estimated 90.4% reduction in cumulative incidence between the ages 35 and 64 with a three year screening interval. We used the model to derive corresponding figures for the reduction in cumulative incidence of lethal invasive cancer (Table VI.1). The IARC statement that the screening interval should be three years or less corresponds with a reduction of at least 90%. If this same percentage is applied to reduction in mortality from cervical cancer our model calculations show that an interval of five years is still acceptable. More frequent screening will give little additional benefit. For example, changing from a five year to a three year interval implies that all efforts (taking smears, cytological analysis, follow up, etc) increase by two thirds, whereas the benefits increase by less than 5%, resulting in a 14 times worse marginal cost effectiveness ratio.

Table VI.I. Percentage reduction in the cumulative rate of cervical cancer incidence and mortality in women aged 35-64 for different frequencies of screening*.

Interval between screenings (years)	% reduction in: cumulative incidence (A)	cumulative incidence (B)	cumulative mortality (C)
1	93.5	95.4	96.9
2	92.5	93.3	95.3
3	90.4	91.0	93.7
4	88.4**	88.3	91.8
5	83.6	85.2	89.7
10	64.1	68.1	80.0

*(A): calculated by the IARC working group, see Table V in the IARC paper (IARC, 1986); (B) and (C) are model based predictions.

**Calculated from Table III in the IARC paper (IARC, 1986)

In an earlier analysis we used a more detailed model based on Dutch epidemiological, medical, and economic data to predict the effects and costs of cervical cancer screening policies (Habbema *et al.*, 1988). For the Netherlands we calculated a cost effectiveness ratio of about £8000 per life year gained for a screening policy with five yearly invitations between the ages of 37 and 62, assuming an average attendance rate of 65% (Habbema *et al.*, 1988, Koopmanschap *et al.*, 1990a). In this model the additional effects of a policy with more frequent invitations (three yearly between ages 36 and 63) are higher than in the case with 100% attendance because of irregular participation patterns. The marginal costs per life year gained with the three yearly policy in comparison with the five yearly policy were about £25 000. Whereas the cost per life year gained of £8000 for a five yearly interval may be considered acceptable, the marginal costs per life year gained by screening every three years are high in comparison with the costs of many other health care facilities. In addition, we found that increasing the screening frequency will worsen the balance between favourable and adverse health effects (unnecessary treatment of women with false positive results and regressive lesions) of screening (van Ballegooijen *et al.*, 1990).

Our analysis is based on published data from studies that are now more than 10 years old. Reports of an increasing incidence and mortality, especially in younger women, do not influence our results since our estimates are predominantly determined by the duration of preinvasive stages and the sensitivity of the screening test. These factors could also change over time (Elliot *et al.*, 1989), stressing the need for continuous re-evaluation of the protection afforded by negative smear test results.

Conclusion

The most serious negative effects of cervical cancer are early death and the serious morbidity associated with advanced disease. Therefore, reduction of the risk of death from cervical cancer should be the primary criterion in evaluating screening policies. Reduction in incidence of invasive cancer carries an additional benefit since some major therapeutic procedures and their associated morbidity will be avoided, but this should be considered in conjunction with the negative health effects of diagnostic and therapeutic procedures that are induced by screening (van Ballegooijen *et al.*, 1990).

The use of reduction in incidence as a proxy for reduction in mortality is appealing. However, we have shown that protection against invasive cancer underestimates protection against mortality. The two criteria will lead to different recommendations: screening intervals based on mortality are about 50% longer than those based on incidence. This difference is caused by the good survival of women with screen detected invasive cancers. On the basis of our calculations of the reduction in cervical cancer mortality we conclude that screening intervals of five years are appropriate. Regular screening at this interval in the age group 35-64 is expected to give about 90% reduction in the risk of dying from cervical cancer. More intensive screening will give little additional benefit, and should be discouraged in view of the adverse effects and the high costs.

VII. DISCUSSION AND CONCLUSION

The main goal of the project "decision making about screening for cancer" has always been to support decision making by predicting the public health effects and other consequences of screening. In making these predictions, it is important that all factors that may influence the effects of screening, and interrelations between these factors, can taken into consideration (Habbema *et al.*, 1978). Some of these factors can be controlled or influenced by the screening policy. For example, persons may be invited at certain ages, and the quality of the screening test may be influenced favourably by adequate training of personnel and by a quality control system. The impact of policy decisions on the effects of screening is far from straightforward and depends on characteristics of the diseases and of the screening tests. An important secondary goal of the project is therefore to acquire a better understanding of these underlying processes on basis of available empirical data. Because of the complicated dynamic relations involved, modelling is indispensable in decision analysis of screening.

In this thesis, which focuses on the secondary goal, I have described the first steps in model-based evaluation of screening for cancer. These steps include:

- a survey of existing knowledge and data and of the main issues in cancer screening (Chapter I, see also Habbema *et.al.*, 1982ab),
- detailed analysis of data from screening projects (see Habbema *et al.*, 1983, 1986)
- development of quantitative modelling tools (Van Oortmarssen *et al.*, 1981, Habbema *et al.*, 1984,1987),
- model-based analysis of data related to breast cancer screening (Chapters II and III) and cervical cancer screening (Chapters IV, V and VI, see also Habbema *et al.*, 1985).

The practical relevance of this work with respect to decision making about screening is expressed in the subsequent cost-effectiveness studies of breast cancer screening and cervical cancer screening, and has been reported in two PhD theses (De Koning *et al.*, 1993, Koopmanschap *et al.*, 1994) and in a number of publications and reports.

I will discuss two topics in this chapter: the methodology of model development and model validation and the applications to breast and cervical cancer screening. Attention will be given to the remaining issues that may be addressed in future model-based analysis of cancer screening.

Model development

The first step in applied modelling of disease control programmes is to identify the key factors that determine the results of such programmes, and to establish the main relations between these factors and the results (Habbema and van Oortmarssen, 1994). This step involves review of the literature, inspection of data from screening programmes, and consultation of experts, and has resulted in definition of a basic structure of an epidemiometric model for cancer screening. This model consists of a number of modules describing the preclinical, screen-detectable stage of the disease (with parameters duration, regression and progression), the treatment (parameters survival and improvement in prognosis as a result of early detection), the screening test (parameters sensitivity and specificity), and the characteristics of a screening policy (age groups to be screened, screening intervals, participation rates). Refinements and extensions to this basic model have been identified, e.g. multiple disease stages, more than one screening test, association between risk and participation in screening.

A general framework has been described (Habbema *et al.*, 1984) which has proven to be applicable to breast cancer screening and to cervical cancer screening. The core of the framework is a single one-dimensional disease history consisting of a number of successive disease states. A model for a specific disease is specified by defining its states and the transitions between states. State definitions are generally chosen on basis of classifications used in available data: the model is data-driven. Transitions are defined by declaring dwelling time distributions in states that are independent of the preceding states and transitions (Semi-Markov property). In reality, staging of the cancer is based on more than one dimension, for example tumour size, node involvement, distant metastases, and diagnosis (yes or no). In the models described in this thesis, and especially in the breast cancer screening models described in Chapters II and III, these dimensions could be combined readily in defining the disease states. However, in further extensions to these models, and in modelling screening for other cancers, for example cancer of the prostate, colorectal cancer, and lung cancer, the one-dimensional Semi-Markov approach may well lead to restrictions because not all dependencies between the variables involved can be adequately translated into the Semi-Markov transitions and durations. For example: lymph node status depends on tumour size, and clinical detection depends on tumour size and lymph node status.

One way out of this problem would be to refrain from the disease stages, and build a model which only considers events, such as start of the preclinical screen-detectable stage, clinical diagnosis, detection at a screening test, death from the disease. The model is specified by describing the functional relations between these events. This type of models has been proposed by David Eddy (1980,1981,1988,1989,1990). The disadvantage of this event-driven approach is that important information in the data is neglected, for example about the differences in stage distribution between screen-detected cases, interval cancers, and clinically diagnosed cancers.

A better alternative would be a disease history which consists of interacting parallel processes such as tumour growth, node involvement, presence of clinical symptoms. This approach, which was used in the breast cancer model of Shwartz (1978ab, 1980, 1981, 1992), may be extended further to model interaction between different disease processes. This would be useful in modelling the interaction between HPV infection and progressive preinvasive cervical cancer, and in modelling the risks of developing malignant colorectal cancer depending on the presence of benign polyps developing in different parts of the colon and that may be detected by different screening tests.

Collaboration with epidemiologists and other experts of screening programs and access to data of such programs has proven to be essential in building well-founded and useful models. Members of our research group participated in the Evaluation committee of the Dutch pilot projects of cervical cancer screening (Evaluation Committee, 1989), in the IARC working group on evaluation of cervical cancer screening programmes (Hakama *et al.*, 1986), in the Dutch evaluation study of the cost and effects of breast cancer screening (Van der Maas, *et al.*, 1989). At present, we are participating in the evaluation teams of the national screening programmes for breast cancer and cervical cancer (LETB, 1992, 1993, 1994). In developing the models, we analyzed data-sets of major screening projects: the statistical data set of the HIP study (Habbema *et al.*, 1983, 1986), the tape of the British Columbia Cohort Study (Van Oortmarssen and Habbema, 1986), and several datasets of the DOM and Nijmegen projects for breast cancer screening and of Dutch projects for cervical cancer screening.

In our implementation of the cancer screening models into computer programs, two approaches emerged: a general and comprehensive microsimulation MISCAN approach, and a simplified numerical-statistical approach which has to be developed from scratch for each specific application.

Initially, much emphasis was put on the requirement of taking all (possibly) relevant factors into consideration in evaluation of screening policies. To meet this requirement, the MISCAN *microsimulation* computer programme has been developed for simulating models that comply to the general framework. The choice of microsimulation was mainly based on previous experience with the *macrosimulation* model SCRMOD developed by E.G. Knox (1973,1975). Being the best comprehensive cancer screening model at that time, it nevertheless turned out that intrinsic properties of macrosimulation seriously limit the number of interrelations that can be included in the models. Both memory usage and simulation time will increase tremendously when new independent variables are added to the model. Microsimulation offers greater freedom for including interactions, and implementation of model extensions to the computer programme is quite easy. Memory usage and simulation time will only increase more or less linear when adding variables to the model. Complex invitation schedules and screening patterns can be incorporated easily in MISCAN. For example, the history of opportunistic and organised cervical cancer screening in The Netherlands, including its long-term effects, has been taken into account in predicting the impact of different future screening policies. And in evaluating breast cancer screening in The Netherlands, a practical complication such as the gradual build-up of the nationwide program has been explicitly included in the model. This flexibility of microsimulation is especially useful in making detailed predictions of the effects and costs of different screening policies in the setting of a real population. All cost-effectiveness calculations have been solely based on MISCAN calculations.

In model-based analysis of screening data, use of (MISCAN) microsimulation enabled detailed simulation of experimental screening projects, for example the randomized breast cancer screening trials in New York (HIP study), and the pilot projects in Utrecht and Nijmegen. An important advantage of microsimulation in analyzing screening data is that the characteristics of the project design and the data can easily be included explicitly in the model. Examples are the age-distribution at entry in a study, different durations of intervals between screenings, definition of cases in randomized trials and in case-control studies, and the precise calculation of person(-year)s at risk. Moreover, output of the model can be produced in the same format that is used in the projects, enabling direct comparisons between model and observed data.

However, use of microsimulation in data analysis has a number of drawbacks that are related to the variability of the simulation output (see also Chapter I). In estimating model parameters from the observed data of a screening project, the standard maximum likelihood approach cannot be applied. Moreover, the estimation procedures will take much computer time because of the relatively long duration of a single simulation run, which needs to be large to reduce the variability of the output), and also because special optimization procedures are required that take random variation in the model output into account, and that are be less efficient than standard optimization procedures for deterministic models.

A statistical-numerical model based on maximum-likelihood methodology will often be worth consideration. The choice between microsimulation (MISCAN) and the statistical-numerical approaches depends on the objectives of the analysis, the characteristics of the data under study, and the amount of time available for the analysis. Screening models used in statistical-numerical data analysis are generally simpler than MISCAN models, either by concentrating on particular aspects of the screening process or by giving a simplified description of the overall model. Modification or extension of the model structure will in general require a substantial effort in deriving mathematical equations and implementation in a computer program - if possible at all. An attractive option is to combine the two approaches, starting with a formal statistical analysis of part of the data with a limited model, and extending the analysis to other data by implementing the full model in MISCAN.

Milestones in the statistical-numerical modelling of cancer screening are the "one stage - one test - one examination" of Zelen and Feinleib (1969), and the "one stage - one test - multiple examinations" model for progressive screen-detectable cancer developed by Walter & Day (1983). The latter model is very useful for a quick combined analysis of detection rates and interval cancers for cohorts with specified screening patterns, yielding estimates of the mean duration of the screen-detectable stage and the sensitivity of the screening test, and has been used as an auxiliary tool in the MISCAN analysis of breast cancer screening data (Chapters II and III). The model of Walter and Day has been extended to two screening tests and two progressive stages (Alexander, 1989), which already requires substantial mathematical and programming efforts.

For a cervical cancer screening model aimed at analyzing detection rates and interval cancers, further refinements are necessary. First, the proportion of (pre-invasive) screen-detectable lesions that will regress spontaneously cannot be neglected when detection rates are included in the analysis. Second, age has to be included explicitly in the model because of the long duration of screen-detectable stages and because of considerable age-dependencies in the model. The resulting models lead to a rather complicated set of equations (see Chapter IV), and the computer implementation is specific for the data available. The data from British Columbia used in Chapter IV are relatively straightforward, describing two birth cohorts that are assumed to have the same basic risk of developing cervical cancer. A similar analysis of Dutch data is complicated by a number of factors: missing data, mix of opportunistic and organized screening, apparent risk differences between cohorts, etc. In view of these complications, a MISCAN approach was chosen for analyzing the Dutch data. One of the starting points was the disease model which resulted from the British Columbia analysis reported in Chapter IV. Other starting points are an Age-Period-Cohort analysis of (pre-screening) cervical cancer mortality data, and estimation of age-dependent survival probabilities on basis of Dutch incidence/mortality ratios and survival data from literature.

This example shows that in modelling screening for cancer, several tools can be combined fruitfully. Microsimulation is then used in the stages in which integrated analyses are being carried out and in making predictions for the effects and costs of screening strategies. But when a microsimulation model like MISCAN is already available, it can also be used in data analysis, especially in exploratory analyses of the data, or in exploring the robustness of the findings of statistical-numerical models.

Validity of the model

A model is only a highly simplified representation of the complex processes involved in mass screening for cancer. The validity of a model is limited by statistical uncertainty, reliability of the data, correctness of the computer programme, misspecification of the model, and uncertainties related to transfer of the model in space and time.

1. The statistical uncertainty caused by the limited number of observations that were used in estimating and testing the model will influence the goodness of fit score, and it will also be reflected in the size of the confidence region for the values of the model parameters. In the analyses presented in Chapters II, III and IV, we used all data available in order to narrow down the size of the confidence regions. However, this implies that no independent data were available for formal validation of the models.

   In using microsimulation, the additional uncertainty caused by the variability of the simulation results had to be taken into account in calculation goodness of fit or confidence regions.
2. Reliability of the data. Available data that are used in quantification of the model may be affected by various sources of bias. In the context of screening for disease, special types of bias occur that have been discussed in Chapter I: lead-time bias, length biased sampling, overdiagnosis bias, and self selection bias. The mechanisms leading to these biases are often automatically included in the model (lead time and

length biased sampling) or can easily be added to the model. Models are very useful in studying their implications, for example lead time and length biased sampling have been investigated with the "one stage - one test" model of Zelen and Feinleib (1969).

More cumbersome are problems caused by the quality of the data, for example caused by incomplete registration, errors in data processing, or changes in classification procedures or in diagnostic or treatment protocols. Sometimes, the latter biases lead to quite drastic inconsistencies that can be detected. For example, a change in the referral policy with respect to cytologic classification "moderate dysplasia" in the Dutch pilot projects of cervical cancer screening caused an sudden increase in detection rates of histologically confirmed abnormalities. And the mortality trends of cervical cancer as reported by the Dutch Cental Bureau of Statistics (CBS) show a dip between 1964 and 1971, that might be interpreted as an early start of the decreasing trend in later years. Further inspection revealed that a similar decrease was found for "uterus corpus" and a considerable increase in mortality classified as "uterus, unspecified", reflecting a temporary change in classification of these cancer sites. Close contact or collaboration with investigators from screening projects is important for correct interpretations in model-based analysis of the data.

3. Reliability of numerical implementation of the model. Of course, the computer programme should be properly validated. Validation of a microsimulation programme can profit from the individual histories that can be checked. But the variability in the simulation outcomes may sometimes lead to problems. More than once we have incorrectly interpreted unexpected outcomes resulting from random variation as caused by errors in the input specification or in the programme code. The use of different approaches (including manual calculations, micro-simulation, and numerical integration methods) is the best (but time-consuming) way for checking the outcomes of a model.

4. Possible misspecification of the model.

   A model which gives a good fit of the data is not necessarily a correct model. For example, certain crucial aspects might be neglected in the model, or incorrect assumptions might be made. These misspecifications may result in erroneous parameter estimates and confidence regions, and incorrect results in extrapolation of the model. For example, predictions about different screening policies based on incorrect estimates of the underlying disease and screening processes will most likely also lead to incorrect recommendations.

Misspecification of a model may be a deliberate choice, for example motivated by lack of time or because lack of sufficient data for testing further refinement of the model. But misspecification can also occur by accident, by overlooking alternative explanations of the data.

The HIP results had been analyzed with two (highly) simplified models. Walter and Day (1983) only considered the duration of preclinical screen-detectable stage and the sensitivity of the screening examination without making a distinction between the outcomes of the two screening tests (mammography and physical examination). Goldberg and Wittes (1978) only considered the test sensitivities on basis of the cross-tabulation of outcomes of the two tests, neglecting the information provided by the interval cancer rate and the trend in detection rates at subsequent tests. The two analyses gave quite different point estimates and confidence intervals of the sensitivity of the screening examination.

Our analysis of the HIP breast cancer screening data (Chapter II) showed that adding more details to a model will narrow down the confidence interval for estimates of the mean duration of preclinical stages and sensitivity of the two screening tests (mammography and physical examination). The estimates for mean duration are longer, and for sensitivity lower than in the Walter & Day model. The confidence region is roughly equal to the intersection of the ones found by Walter & Day and by Goldberg & Wittes.

Age-specific comparisons between the model and HIP results suggested age-dependency of sensitivity and/or of duration of preclinical stages, but a model with age-dependent duration did not give a significant improvement of the fit, and we decided to assume a constant duration. This is an example of misspecification caused by lack of data, since all analyses of more recent datasets indicated a significant increase of the mean duration with age, see for example Chapter III.

The possibility of serious misspecification can never be ruled out, but it can be minimized by performing extensive robustness analyses regarding the parameter estimation (see the *discussion* section of the cervical cancer model presented in Chapter IV). A potential remaining source of misspecification in the cervical cancer model is discussed in the *Application* section of the present chapter.

5. Uncertainty because of transfer in space and/or time.

This problem relates to analysis of different screening datasets with a single model or with models which have a number of assumptions in common. The additional problem of transferring model assumptions in time is especially relevant when models that have been fitted to local data are being used for making predictions.

In the cancer screening model, a distinction is made between the "general" biological factors and "local" factors that are influenced by the specific conditions in a region or country. However, the extent of variability between areas is not known precisely, and factors that are assumed to be biological or otherwise invariable may in reality be differ between areas or vary with time. For example, in the analysis of risk reduction after negative Pap smears (Chapter V) we assume that the relative risk of participants in cervical cancer screening, estimated to be 0.74 in British Columbia, applies to the other screening projects as well. But it is likely that the distribution of risk of cervical cancer, and its association with participation, differs considerably between countries.

The MISCAN models have been repeatedly transferred between screening projects. In such a transfer, model specifications concerning local characteristics of the projects (for example: the level of incidence and mortality, the screening policy, the participation rate) are adapted, whilst the parameters of the underlying processes are preserved. For example, the breast cancer screening model was first constructed on basis of the HIP data, and then transferred and modernized on basis of data from the Utrecht and Nijmegen projects. The resulting model was transferred to the Swedish trials in Kopparberg/Östergötland and Malmö for estimation of the effect of early detection and treatment on mortality, and finally to the national situation in the Netherlands for making cost-effectiveness predictions of screening policies. This breast cancer model has been transferred and adapted to the local characteristics of other countries (e.g., Australia, Italy, Germany) for evaluating cost-effectiveness of breast cancer screening (Carter *et al.*, 1993, Van Ineveld *et al.*, 1993, Paci *et al.*(in press), Beemsterboer *et al.*, 1994).

Similarly, the cervical cancer model based on data from British Columbia (Chapter IV) has been transferred to the situation in the Netherlands, and was extended and calibrated to the specific Dutch circumstances. This model has also been used to make future predictions about cost-effectiveness of screening in The Netherlands (Van Ballegooijen *et al.*, 1993) and in Italy.

The advantage of modelling is that the potential biases resulting from uncertainty about local or future parameter values can be addressed by performing sensitivity and uncertainty analyses. An example of the potential usefulness of models is in addressing problems of transferring observed screening effects is in meta-analyses. Several meta-analyses of

RCT's for breast cancer screening have been published, but in all analyses the overt differences between the studies are neglected. Models can be used to control for the local characteristics, and investigate the remaining variability between projects.

Further work

The limitations posed by the assumption of a one-dimensional semi-Markov disease process are beginning to cause problems in implementing extensions to the breast and cervical cancer models, and will presumably also lead to problems in modelling colorectal cancer screening. A more general framework in which a number of interdependent processes can be modelled deserves serious consideration. This could lead to a new general-purpose version of MISCAN. An alternative would be to build special models for each disease on basis of a common library of subroutines.

Complete revision of the MISCAN programme has another potential advantage. In the past, we have been asked regularly for a copy of MISCAN. For a number of reasons, we have always been reluctant in distributing MISCAN. An important reason in this respect is that we developed as a tool for our own research activities, which means that it includes a number of peculiarities, that some options that had been built in (but not used) have never been tested and checked, and that only crude documentation is available. In revising MISCAN, it could be tried to develop a core version for distribution. However, this would require a considerable additional effort, and it will probably be difficult to get necessary funding.

Two natural extensions of our microsimulation work are currently being considered.

First, microsimulation can be used to generate fictitious data sets on basis of assumptions about effects of screening, which can be used to investigate study designs and methods for analyzing data. At the moment, a project is going on in collaboration with the National Cancer Institute (USA) to study the methodology of case-control studies for analyzing the effects of screening for breast cancer.

The second extension of microsimulation is being developed in collaboration with the International Institute for Applied Systems Analysis (IIASA). The objective is to develop and implement a method for optimization of screening strategies, i.e. finding strategies that give the best balance between costs and effects for a given number of screening invitations. Highly simplified models can be optimized by writing down the equations and using conventional optimization techniques. Models that take the principal factors into account, such as the cervical cancer screening model described in Chapter IV, are already too complicated for these techniques. Models of medium complexity can optimized with methods that do not require computation of derivatives, such as the downhill simplex method, see for example Press *et al.* (1988). This method was used for optimization of cervical cancer screening policies by Gustafsson and Adami (1992) and by Verhoeven (1994). A promising alternative for stochastic simulation models is the Stochastic Quasi-Gradient (SQG) technique, which can be regarded as a special form of microsimulation. Starting from a given screening strategy, the strategy is iteratively adjusted after each simulated life history. After a large number of histories, the resulting strategy is expected to be very close to the real optimum (as in the case of normal microsimulation, the outcome is an estimate of the true value). The SQG method requires strict regularity conditions, and cancer screening models require considerable reformulation, see Oortmarssen & Ermoliev (1994). The main advantage of SQG optimization is that it is very fast when compared to alternative optimization methods.

It remains to be seen whether the SQG approach can also be applied to comprehensive models for breast cancer and cervical cancer that are used for cost-effectiveness calculations, and whether it can be applied in estimating values of the model parameters from screening data. An alternative option is to use general optimization routines for stochastic models, that operate as a shell around MISCAN and can both be used for data analysis and policy optimization, without the need for reformulating the disease/screening model. A computer package MINIMIZE containing five different optimization methods for stochastic models is under development at our department and is currently being tested for various model applications.

Application to screening for cancer

Breast cancer

In discussing misspecification errors in the previous section, I explained that our model-based analysis of the HIP study (Chapter II) can be regarded as an integration of two simplified models. Additional refinements in our HIP model include a distinction between two tumour size categories which differ in level of sensitivity, and the introduction of age as one of the variables of the model. The screening results from the Dutch pilot project in Nijmegen, where all women were invited from age 35 onward, allowed for analyzing the age-dependency of preclinical dwelling times which was suspected but could not be demonstrated from the HIP data. A good fit of the data from Nijmegen was obtained by assuming increasing mean duration with age (from 1.6 years at age 40 to 4.7 years at age 70). The improvement in mammographic techniques is reflected in much higher estimates for the sensitivity in comparison with the HIP analysis. Most data from the other Dutch pilot project in Utrecht, which had an entirely different invitation scheme, could be fitted with the same assumptions about preclinical disease and test sensitivity.

The breast cancer model has been extended to include clinical aspects (diagnostic assessment and treatment), economic aspects (cost of screening, diagnosis and treatment), and quality of life information. The resulting *cost-effectiveness models* have been used for predicting the effects and costs of different screening policies in the Netherlands (Van der Maas *et al.*, 1989, De Koning *et al.*, 1991).

In recent years, the breast cancer model has been refined further in comparison with the description given in Chapter III. The subdivision in tumour diameter categories of invasive cancers has been changed from <10, 10-19, 20+ mm. into ≤5, 6-10, 11-20, >20 mm. in order to be compatible with the TNM staging system, and a further subdivision has been made into lymph node negative and positive stages, using an ad hoc extension to MISCAN. Other refinements include a distinction in the test-sensitivity and improvement in prognosis in women under and above age 50, and a revision of age-dependent duration of preclinical stages at old ages (65+). For details see Chapter 7 in (De Koning *et al.*, 1990b), and Chapter 6 in the third evaluation report of the National Evaluation Team (LETB, 1994). The model has been transferred to the situation in Germany, and this German version has been fitted to available incidence, mortality, and screening data (Beemsterboer *et al.*, 1994). Similar studies of the costs and effects of screening have been done for Australia and Florence (Italy). Recently we started another evaluation of screening in Catalonia (Spain).

The present model, which is in good agreement with the data from the Dutch pilot studies regarding the detection rates and interval cancers, has been used to make detailed predictions of detection rates, interval cancer rates, and stage distribution in the Dutch national screening programme. Thus far, the results of the programme are in agreement with these predictions (LETB, 1992, 1993, 1994), with one exception: the models predicted that the stage distribution at a repeat screening is better than in the first round, but in practice the distributions are similar. Two possible explanations are false reassurance - negative screening results might lead to disregarding of early symptoms - and a more complex natural history than assumed in the model, with fast growing cancer spending relatively much time in advanced stages.

The model is roughly compatible with the Swedish randomized trials regarding the reduction in breast cancer mortality: it predicts the average mortality reduction of these trials when allowing for the specific characteristics of the trials (age-groups, screening intervals, duration of follow-up). The improvement in prognosis following early detection is a crucial parameter of the model, and further analysis of the Swedish trials including the detection rates, interval cancer rates, and stage distributions, would contribute to the credibility of the model.

The most important question in breast cancer remains at present the reduction in mortality resulting from screening women below age 50, which we addressed on basis of the HIP data (Habbema *et al.*, 1986). The debate concerning interpretation of the HIP results is still continuing, but is of limited relevance for present day decision making because of the major improvement in mammographic techniques since the late 1960's, and the marked trend towards a more favourable stage distribution of clinically diagnosed cancers.

On basis of the favourable HIP findings for women below age 50, one could have expected a clear reduction in the modern trials in view of the improvement in technique. However, after 6-12 years of follow-up, the trials in Sweden only show a non-significant 10% reduction in mortality in this age-group (Nyström *et al.*, 1993), and the Edinburgh trial also shows a small non-significant reduction at 10 years of follow-up (UK trial, 1988, Roberts *et al.*, 1990, Fletcher *et al.*, 1993). In the 50+ group, a significant mortality reduction occurs in some of the separate Swedish projects and also in the pooled Swedish results (Nyström *et al.*, 1993). A smaller reduction in mortality in younger women may be explained by the shorter average duration of pre-clinical stages, a lower sensitivity of mammography, and the more favourable prognosis of clinically diagnosed cancers (Adami *et al.*, 1986). Note however that a considerable proportion of the breast cancers detected at repeat screenings in the age group <50 *at entry* will have been diagnosed at ages over 50 years. Indeed, MISCAN simulations of the pooled Swedish projects, in which we assumed no effect for cases diagnosed below age 50 and a constant level of mortality reduction after this age, predicted a mortality reduction of almost 10% in the age group <50 at entry.

It is still possible that the history of the HIP study, in which the initially small difference in younger women only became significant in the long run, will repeat itself. In the mean time, it is hard to justify mass screening for women below age 50 in view of the uncertainty about favourable effects, and the inevitable adverse effects of which some are relatively important at younger ages (false positives, radiation risk). A more appropriate use of additional resources would be to intensify screening above age 50. Indeed, one may argue that there should be a moratorium on all mammography for symptom-free women below age 50 outside randomized controlled trials (Lancet, 1991). New randomized trials in the age group 40-49 have started in the UK (Vessey, 1991), and may be followed by other countries.

Another important question concerns the effectiveness and the balance between favourable and adverse effects and costs of screening above age 70 (Boer *et al.*, submitted for publication). A related issue is the potential increase in "opportunistic" screening in these age-groups as a result of frequent organized screening in the 50-69 age-group, and the hazards of radiation in young women. Improvements in therapeutic possibilities and an autonomous trend towards earlier clinical diagnosis might reduce the favourable effects of screening in all age groups.

Cervical cancer

Until recently, no complete model of cervical cancer screening existed that had been carefully tested against available data. See Prorok (1986) for an overview of models for cervical cancer screening. Modelling work by Knox (1973), Coppleson and Brown (1975), and Parkin (1985,1986) suggested a long mean duration and considerable amount of regression of pre-invasive stages, but did not include formal parameter estimation or testing of hypotheses. Brookmeyer and Day (1987) presented a simplified model and estimated parameters related to detection of progressive pre-clinical stages, but did not use a substantial part of available data: detection rates of pre-invasive cancer.

The simplified model that I developed can be regarded as a generalisation of the latter model. All major factors playing a role in cervical cancer screening are covered by the 10 parameters of this model, and it gives a good fit of detailed data from British Columbia. The estimated parameter values indicate that regression is especially important in lesions which develop in young women. This supports the conclusions from a conventional epidemiological analysis of the British Columbia screening data by Boyes *et al.*, 1982. The surplus value of the model is that it also gives estimates for the duration (mean and variability) of pre-invasive lesions and for the sensitivity of the Pap smear, which are crucial parameters in comparing screening policies. An important presupposition of the model is the use of a single distribution for the duration of both progressive and regressive pre-invasive cancers, which is also independent of age.

Similar estimates, based on modelling Swedish cervical cancer screening data, were published by Gustafsson and Adami (1989, 1990).

The model (including parameter values derived from the British Columbia data) has been used to analyze the findings of the IARC collaborative study concerning reduction in risk of invasive cancer for women who have had one or more negative smears (Chapter V). This analysis shows that the risk reduction reported by the IARC group might be too pessimistic, since it includes invasive cancers that are detected by screening. The model is used to calculate the risk reduction for clinically diagnosed cancers, using assumptions about the frequencies of further smears and the duration of the preclinical invasive stage of cervical cancer. It appears that when the influence of screen-detected invasive cancers is taken account, good agreement exist between model assumptions about the duration of preinvasive stages and the findings of the IARC study. These findings are used to challenge the recommendation from the IARC study group that the screening interval should be three years or less (Chapter VI). This recommendation was based on the risk reduction for invasive cancers, which means that the favourable effect of early detection of (micro-) invasive cancers is ignored. In our view, reduction of cervical cancer mortality should be the criterion for evaluating the effect of screening. Screening intervals of 4 or 5 years will already give a 90% reduction in mortality in the age groups involved. More frequent screening will lead to a high marginal cost-effectiveness ratio, and also give lead to an increase of the adverse effects of screening: treatment of women with false positive test results and regressive lesions.

A refined version of the simplified model was implemented in MISCAN, and has been used in analyzing the British Columbia data (yielding similar results) and screening data from the Netherlands. Detailed analysis of the Dutch data from the pilot regions (1976-1985) was hampered by many problems: clear risk differences exist between birth cohorts, Pap smears are being made both within and outside the organized programme, classification of smear results and follow-up policies in case of abnormalities were changed in the course of the pilot studies, etc. Still, the parameter values derived in the British Columbia analyses gave a reasonable overall fit of the Dutch data (see Habbema *et al.* 1988: annex C).

This MISCAN cervical cancer screening model has been used for calculating the costs and (favourable and adverse) effects of screening strategies (Van Ballegooijen *et al.*, 1990, Koopmanschap *et al.*, 1990a). The conclusion of this evaluation was that screening should be extended to both younger and older age groups. It was felt that a further analysis of data for age groups (35- and 55+) was indicated. This analysis resulted in small adjustments to the model, but reinforced the conclusions about screening strategies (Van Ballegooijen *et al.*, 1993).

At present, the most important issue is the Human Papilloma Virus (HPV)-based screening test. Potentially, introduction of this test could reduce the number of screening examinations in the population, and possibly also discriminate between progressive and regressive preinvasive lesions. HPV stages and the HPV test have been added to the model, which will be quantified on basis of available data. Evaluation of screening policies is now complicated because of the possibilities to use either of two tests (Pap or HPV) or the combination. Another issue that deserves continued attention is balance between adverse and favourable effects in detection and treatment of early lesions (CIN I/II).

Modelling of breast and cervical cancer will continue as part of the evaluation of nationwide screening programmes in The Netherlands. Complementary studies are being carried out of breast cancer screening in Catalonia, and of the HPV-based screening for cervical cancer. Other model-based evaluation studies address prostate cancer screening, and genetic screening. Modelling of colorectal cancer screening is under serious consideration.

Conclusion

Model-based analysis of some of the best available datasets of screening has resulted in a better understanding of natural history of breast and cervical cancer and of the impact of screening. Our analyses of breast cancer screening data have shown that detection rates, interval cancer rates, and stage distributions of different projects can be fitted reasonably well with a fairly simple model, with age-dependent duration of preclinical screen-detectable stages. Analysis of cervical cancer screening data strongly supports the existence of regressive lesions, and resulted in estimates for the duration of preinvasive lesions. A reasonable explanation of complicated epidemiological trends in The Netherlands during the past decades was obtained.

Model-based analyses of screening projects is one of the building blocks for prospective evaluation of costs and effects of screening. The outcomes of evaluation of breast cancer screening have had a noticeable influence on the decision to start nationwide screening in The Netherlands. And the extensive comparison of screening policies for cervical cancer has contributed to the recent decision to change from a 3-years to a 5-years screening interval. The main objective of the project, supporting decision making about screening for cancer, has been realized.

SAMENVATTING

Achtergrond van de studie

In Nederland vindt al sinds 1977 op landelijke schaal bevolkingsonderzoek op baarmoederhalskanker plaats, en is sinds kort begonnen met een landelijk bevolkingsonderzoek op borstkanker. Deze twee kankers vormen voor vrouwen tussen 30 en 65 jaar een belangrijke doodsoorzaak gezien het feit dat ze rond 20% van de totale sterfte in deze leeftijdsgroep veroorzaken. Het doel van deze bevolkingsonderzoeken, waarbij jaarlijks enige honderdduizenden vrouwen uitgenodigd worden om zich te laten onderzoeken ("*screenen*"), is om de kankers in een vroeg stadium op te sporen en te behandelen. Hierdoor wordt volgens de thans beschikbare kennis de kans om aan een van deze ziektes te overlijden aanzienlijk verminderd.

Aan bevolkingsonderzoek zijn echter naast de verminderde kans om aan kanker te overlijden ook andere effecten verbonden, waaronder een aantal ongunstige. Voorbeelden van ongunstige effecten zijn de ongerustheid die het bevolkingsonderzoek veroorzaakt bij de uitgenodigde vrouwen (vooral in de tijd tussen een uitnodiging en het bekend worden van de uitslag), en het ontdekken en in sommige gevallen behandelen van afwijkingen die nooit zouden zijn uitgegroeid tot een kanker die aan de hand van klachten wordt gediagnostiseerd. Hiertegenover staan echter ook gunstige neveneffecten: veelal zal na vroege opsporing met een (veel) minder ingrijpende behandeling kunnen worden volstaan, en in samenhang met de verminderde kans op sterfte aan de kanker neemt ook het risico af op recurrentie en uitzaaiing na de primaire behandeling.

Organisatie van bevolkingsonderzoek is een zaak van de overheid, kankercentra en ziekenfondsen. Bij de beslissing om bevolkingsonderzoek te starten of te continueren spelen naast de verwachting over de effecten ook de kosten een belangrijke rol. Ter illustratie: de jaarlijkse kosten van een bevolkingsonderzoek op baarmoederhalskanker zijn vele malen hoger dan de totale kosten van diagnostiek en behandeling van baarmoederhalskanker patiënten.

Evaluatie van bevolkingsonderzoek

Bij besluitvorming over bevolkingsonderzoek zijn onder meer de volgende vragen aan de orde:

1 zijn de gunstige effecten met voldoende zekerheid aangetoond?
2 wat is de verwachte verhouding tussen gunstige effecten, ongunstige effecten en kosten?
3 welke leeftijdsgroepen dienen te worden uitgenodigd, en hoe vaak?

De eerste vraag kan alleen worden beantwoord aan de hand van bevindingen van experimentele bevolkingsonderzoeken. De ideale opzet is een gerandomiseerde proefstudie (randomized controlled trial, afgekort RCT). In een RCT wordt een welomschreven populatie vrouwen via aselecte keuze ingedeeld in twee groepen die wél of niét worden uitgenodigd. Vervolgens wordt de sterfte in beide groepen vergeleken. Om een duidelijk verschil in sterfte te kunnen aantonen hebben deze proefstudies een zeer grote omvang nodig (tienduizenden of honderdduizenden uitgenodigde vrouwen) en een lange looptijd (meestal wordt pas na 6-8 jaar een verschil in sterfte verwacht).

Antwoorden op de vragen 2 en 3 kunnen niet rechtstreeks uit de resultaten van proefstudies worden afgeleid. Ten eerste zal men deze resultaten willen aanpassen in verband met verschillen in omstandigheden ten opzichte van de proefstudies: verschillen in het voorkomen van de kanker en de sterftekans, verschillen in de deelnamegraad aan bevolkingsonderzoek, etcetera. Ten tweede zal men alternatieven willen overwegen voor in de proefstudies gebruikte oproepschema's (leeftijden, intervallen). Ten derde zal het vaak zo zijn dat de voortschrijdende medisch-technische ontwikkeling aanpassing nodig maakt, immers de resultaten van proefstudies weerspiegelen de situatie van zo'n 10 jaar voordien.

Gezien het ingewikkelde samenspel van ziektebeloop, screeningsonderzoeken, en verandering van overlevingskansen zijn de consequenties van deze aanpassingen alleen met behulp van modellen te berekenen. De constructie van dergelijke modellen voor bevolkingsonderzoek op borstkanker en baarmoederhalskanker is het onderwerp van dit proefschrift. Dit omvat het verzamelen van gegevens, het opstellen van het model, het schatten van de parameters van het model, en zo mogelijk validatie van het model aan de hand van gegevens die niet bij de schatting zijn gebruikt.

Modellering van bevolkingsonderzoek

De constructie van de modellen ten behoeve van evaluatie van bevolkingsonderzoek is een onderdeel van het project "Besluitvorming over Bevolkingsonderzoek" wat in 1977 op het Instituut Maatschappelijke Gezondheidszorg is gestart. Binnen dit project heb ik meegewerkt aan de ontwikkeling van het computersimulatieprogramma MISCAN waarmee vroege opsporing van kanker kan worden doorgerekend. De met behulp van MISCAN geconstrueerde modellen zijn vervolgens gebruikt bij kosten-effectiviteits-studies van bevolkingsonderzoek op borstkanker en baarmoederhalskanker.

Bevolkingsonderzoek op borstkanker

Voor het bevolkingsonderzoek op borstkanker is allereerst een analyse gemaakt van de eerste gerandomiseerde proefstudie, het HIP project[1], dat begon in 1964. De resultaten van het HIP project zijn met behulp van MISCAN geanalyseerd (Hoofdstuk II). Hierbij is een model voor borstkankerscreening ontwikkeld dat goed in overeenstemming is met de resultaten van het HIP project met betrekking tot via screening (mammografie en/of lichamelijk onderzoek) ontdekte kankers, intervalkankers en sterftedaling. De beste overeenkomst werd bereikt via een model met een gemiddelde duur van 2.2 jaar voor de preklinische stadia, en een gemiddelde sensitiviteit van 68% voor de combinatie van de twee screeningsmethoden. Hierbij moet worden aangetekend dat deze waarden gelden voor de stand van zaken in de tweede helft van de jaren zestig. Twee factoren zijn sinds die tijd aanzienlijk veranderd: de stadiumverdeling van via klachten ontdekte borstkankers is gunstiger geworden door de grotere opmerkzaamheid van vrouwen en artsen ten aanzien van symptomen, en de techniek van de mammografie is sterk verbeterd.

Deze veranderingen kwamen naar voren in de MISCAN analyse van de proefstudies voor bevolkingsonderzoek op borstkanker in Nijmegen en Utrecht (Hoofdstuk III). De veranderde stadiumverdeling blijkt uit de gegevens over borstkankers die zijn aangetroffen bij niet uitgenodigde vrouwen. De verbetering van de mammografie komt tot uiting in het aandeel van kleine tumoren (diameter kleiner dan 10 mm.) in de door screening ontdekte kankers. De MISCAN analyse wijst op een toename van 30% voor de gemiddelde duur van de stadia waarin invasieve borstkankers via screening te ontdekken zijn.

[1] HIP: *Health Insurance Plan of Greater New York*, een verzekeringsmaatschappij voor ziektekosten.

Aangezien volgens onze analyse ook de sensitiviteit van de screeningstest aanzienlijk is toegenomen, is de kans op vroegtijdige ontdekking bij een screeningsprogramma op borstkanker sterk vergroot. Dit geldt met name voor oudere vrouwen, omdat we bij stijgende leeftijd een toename vonden van de gemiddelde duur van de vroege stadia van borstkanker: van minder dan 2 jaar op leeftijd 40 tot ongeveer 5 jaar op leeftijd 70.

Het model wist de resultaten van de proefprojecten in Nijmegen en Utrecht goed te reproduceren, met uitzondering van de percentages bij screening ontdekte borstkankers in het eerste deelproject (DOM-I) in de stad Utrecht. Voor de verschillen tussen model en werkelijkheid die hier optreden ontbreekt een afdoende verklaring.

Ook de uit vroege opsporing bij bevolkingsonderzoek voortvloeiende reductie van de kans om aan borstkanker te overlijden is gevoelig voor de ontwikkelingen die sinds het HIP project hebben plaatsgevonden: een betere klinische stadiumverdeling vermindert het potentiële effect van bevolkingsonderzoek, terwijl daarentegen de verbeterde mammografie via nog vroegere opsporing het effect zou vergroten. De effectiviteit is opnieuw geanalyseerd op basis van uitkomsten van de Zweedse gerandomiseerde proefstudies in Kopparberg/Östergötland en Malmö. Allereerst bleek dat het model dat was gebaseerd op de gegevens uit Nijmegen en Utrecht ook de resultaten betreffende via screening ontdekte kankers en intervalkankers uit deze Zweedse projecten redelijk goed wist te reproduceren. Vervolgens is de prognoseverbetering als gevolg van vroege ontdekking geschat.

Bevolkingsonderzoek op baarmoederhalskanker

Bij screening op baarmoederhalskanker via het uitstrijkje wordt getracht niet-invasieve voorstadia van deze kanker op te sporen. Deze voorstadia kennen een lange gemiddelde duur. Wij hebben een aantal analyses verricht op de gegevens uit British Columbia (Canada). Deze gegevens zijn in zoverre uniek dat ze een complete registratie omvatten gedurende de eerste 20 jaren dat in die provincie screening plaatsvond (1949-1969). Deze lange periode maakt het in principe mogelijk om gefundeerde schattingen te maken over de duur van de voorstadia. Hierbij is opnieuw gebruikt gemaakt van het MISCAN simulatiepakket.

Tevens is een speciaal gesimplificeerd model ontwikkeld wat het mogelijk maakt om standaard statistische technieken (maximum-likelihood schatting, likelihood ratio test) te gebruiken voor het schatten van modelparameters en het toetsen van hypotheses over het natuurlijk beloop van de niet-invasieve voorstadia van baarmoederhalskanker (Hoofdstuk IV). Volgens dit gesimplificeerde model zal met name voor bij jonge vrouwen op te sporen voorstadia gelden dat een aanzienlijk deel (±84%) weer spontaan zou verdwijnen (in regressie gaan). Bij oudere vrouwen (de grens ligt in dit gesimplificeerde model bij een leeftijd van 34 jaar) is in veel mindere mate sprake van regressie. De gemiddelde duur van de niet-invasieve stadia is gesçhat op 12 jaar. Tussen het ontstaan van invasieve kanker en de ontdekking ervan aan de hand van symptomen verloopt gemiddeld zo'n 4 à 5 jaar. Met deze aannames worden de resultaten die in British Columbia werden waargenomen uitstekend door het model gereproduceerd.

Een werkgroep van het IARC[2] (waarin wij participeerden) heeft een gecombineerde analyse uitgevoerd van gegevens uit 8 landen over bevolkingsonderzoek op baarmoederhalskanker. Er werden schattingen gemaakt over de afname in het risico op invasieve baarmoederhalskanker die volgt op één of meerdere uitstrijkjes met een negatieve uitslag. Wij hebben het gesimplificeerde model gebruikt om deze "relatieve beschermingsgraad" te berekenen. Hierbij bleek allereerst dat de resultaten van de IARC analyse beïnvloed worden door invasieve kankers die via een uitstrijkje zijn ontdekt. Alleen als hiermee in de modelberekeningen rekening wordt gehouden geeft het model resultaten die goed overeenkomen met de uitkomsten van de IARC analyse (Hoofdstuk V).

Hoewel de kans om aan baarmoederhalskanker te overlijden bij de via bevolkingsonderzoek ontdekte invasieve kankers niet is uitgesloten, is deze toch duidelijk geringer dan bij via klachten ontdekte kankers. Daarom zal het effect van bevolkingsonderzoek op de incidentie van invasieve kankers minder groot zijn dan het effect op de sterfte ten gevolge van baarmoederhalskanker. Wanneer bij vrouwen tussen 35 en 64 jaar iedere 3 jaar een uitstrijkje wordt gemaakt, dan vermindert volgens het model - in overeenstemming met de IARC uitkomsten - het risico op een invasieve baarmoederhalskanker tot 10% van het risico bij niet-gescreende vrouwen. De sterfte aan baarmoederhalskanker reduceert volgens ons model sterker, en wel tot 6%. Bij 5-jaarlijkse uitstrijkjes zijn de percentages respectievelijk 16% en 10% (Hoofdstuk VI).

Aan een uitstrijkje zijn ook negatieve effecten verbonden, zoals het gegeven dat een deel van de ontdekte voorstadia vanzelf zou verdwijnen en dus "onnodig" behandeld wordt. Bij een kort screeningsinterval nemen deze ongunstige effecten veel sterker toe dan de gunstige effecten. Hetzelfde geldt voor kosten die aan het maken van uitstrijkjes zijn verbonden: bevolkings-

[2] IARC: *International Agency for Research on Cancer*, een internationaal instituut voor kankeronderzoek in Lyon.

onderzoek is bij een 3-jaars interval grofweg 66% duurder dan bij een 5-jaars interval. In dit licht bezien valt de aanbeveling van de IARC werkgroep dat minstens éénmaal in de 3 jaar een uitstrijkje dient te worden gemaakt moeilijk te rechtvaardigen.

Evaluatie van de effecten en de kosten van bevolkingsonderzoek

De in dit proefschrift besproken modellen voor bevolkingsonderzoek op borstkanker en baarmoederhalskanker hebben gediend als rekenmodellen bij uitvoerige evaluatiestudies naar de kosten en effecten van deze bevolkingsonderzoeken.

De uitkomsten voor bevolkingsonderzoek op borstkanker bevestigen de keuze om vrouwen tussen leeftijd 50 en 70 iedere 2 jaar uit te nodigen. Naar verwachting zal dit op lange termijn de jaarlijkse sterfte aan borstkanker in Nederland met 650 verminderen. De kosten-effectiviteit wordt geschat op 7600 gulden per gewonnen levensjaar, dit is gunstiger dan voor veel andere medische voorzieningen waarvoor dit bepaald is.

De bevindingen voor het bevolkingsonderzoek op baarmoederhalskanker zijn minder positief over de huidige gang van zaken waarbij een flink deel van de uitstrijkjes bij vrij jonge vrouwen wordt gemaakt, vaak met een korte tussenpoos. Volgens onze berekeningen zou bevolkingsonderzoek het best kunnen beginnen rond de leeftijd van 30 jaar, waarna vrouwen met een niet te kort interval worden uitgenodigd voor een uitstrijkje. Deze uitnodigingen zouden ook na het 55^{e} jaar nog moeten doorgaan. In het meest gunstige geval zouden de kosten per gewonnen levensjaar ongeveer 24000 gulden bedragen, in de huidige praktijk is dit bedrag veel hoger.

Conclusie

In Hoofdstuk VII wordt de balans opgemaakt. Er is een algemeen model voor bevolkingsonderzoek op kanker opgesteld, wat vele mogelijkheden tot variatie biedt, onder meer voor wat betreft de beschrijving van het ziekteproces, en de wijze waarop uitnodiging plaatsvindt. Dit model is geïmplementeerd in het computerprogramma MISCAN. Hiernaast zijn voor specifieke toepassingen gesimplificeerde modellen ontwikkeld en in computerprogramma's vertaald. Beide typen modellen zijn toegepast voor het analyseren van resultaten van gerenommeerde projecten voor vroege opsporing van baarmoederhalskanker en borstkanker, waarbij schattingen zijn verkregen over de belangrijkste parameters van het ziekteproces en de screeningstest. Dit heeft geleid tot een beter inzicht in het ziektebeloop en de mogelijkheden tot vroege opsporing.

De modellen, en dan met name het model voor borstkankerscreening, zijn enkele malen met succes overgeheveld tussen screeningsprojecten, waarbij, na aanpassing aan de verschillende omstandigheden bij deze projecten, de beschrijving van de medisch-biologische processen grotendeels intact konden worden gelaten.

Uiteindelijk zijn de epidemiometrische modellen uitgebreid tot kosten-effecten modellen en gebruikt om voorspellingen te maken over de kosten en effecten van bevolkingsonderzoek in Nederland. De conclusies van deze kosten-effectiviteits studies hebben aanzienlijke invloed op het beleid gehad. Het besluit tot invoering van het landelijke bevolkingsonderzoek op borstkanker is mede genomen op de door ons gemaakte berekeningen over de te verwachten effecten en kosten. En recentelijk is mede op basis van onze aanbevelingen besloten om het oproepschema voor het bevolkingsonderzoek op baarmoederhalskanker te wijzigen en in plaats van een 3-jaars een 5-jaars interval tussen uitnodigingen te hanteren, en de leeftijdsrange te verruimen van 35-53 tot 30-60 jaar.

Ook internationaal bestaat belangstelling voor onze modelmatige evaluatie van bevolkingsonderzoek. In samenwerking met groepen uit Australië, Italië, Duitsland en Spanje zijn of worden kosten-effecten berekeningen gemaakt, waarbij de modellen worden aangepast aan de specifieke omstandigheden in deze landen.

Er valt nog veel te doen aan de verdere ontwikkeling en toepassing van modellen voor bevolkingsonderzoek. Voor een flink deel hangt dit echter af van het beschikbaar komen van nieuwe gegevens en uitkomsten van proefstudies. Dit geldt allereerst voor de vraag naar de effectiviteit van vroege opsporing van borstkanker bij vrouwen jonger dan 50 jaar. Hoewel wij als een der eersten betoogden dat de resultaten van het HIP project ook voor deze groep een gunstig effect lieten zien, kon deze bevinding in latere proefstudies nog niet worden bevestigd. Bij baarmoederhalskanker is juist de vraag of screening voortgezet moet worden na het 55^{e} jaar.

Een nieuwe stap in de modeltoepassing is de bewaking van het verloop van de Nederlandse bevolkingsonderzoeken op basis van de door ons gemaakte voorspellingen. Als de resultaten van het bevolkingsonderzoek hiervan sterk afwijken kan de oorzaak voor het verschil, eventueel met behulp van het model, worden opgespoord. Indien nodig zal het model worden aangepast of kan het model worden gebruikt om aanbevelingen te doen over aanpassingen van het beleid.

SUMMARY

Interpretation of the results of screening programmes and answering the questions involved in planning of mass screening programmes is difficult because of the many factors that determine the direct results of screening, the long-term favourable and adverse effects, and the costs and potential savings. Different methods can be used and combined to address these issues, including experimental studies, careful analysis of existing data, and modelling.

This thesis focuses on development and application of models for interpretation of data from screening programmes for breast cancer and cervical cancer. The work reported is part of a larger project "decision making in mass screening for cancer", in which the models have been applied for predicting the effects and costs of screening policies, using the MISCAN microsimulation programme.

At the heart of models for cancer screening are the *underlying processes* that are crucial determinants of the effects of screening but cannot be observed directly: the preclinical course of the cancer, the sensitivity of the screening test, and the improvement in prognosis following early detection and treatment. In model-based analysis of screening results, parameters describing these processes are estimated and hypotheses about the processes are tested. Two different techniques have been applied: MISCAN microsimulation and a statistical/numerical approach. Advantages of using MISCAN are that identical models can be used for data analysis and for prospective evaluation of screening policies, and that models can be quite detailed. The statistical/numerical approach can only be used with relatively simple models, but has the advantage that it is based on standard maximum-likelihood methodology. Given the availability of MISCAN, implementation of MISCAN-based model will be faster but the calculations will take much more time compared to a statistical/numerical equivalent.

The methodology of modelling screening for cancer is continues to be one of the research issues. Methods for optimization of microsimulation models are being investigated and tested in collaboration with the International Institute for Applied Systems Analysis. MISCAN is being used for generating datasets that are utilized to check methods for analyzing case-control studies of the effect of screening. And we are planning a complete revision of the MISCAN package, largely motivated by difficulties encountered in further refinement of existing models, due to limitations of the current MISCAN version.

The MISCAN simulation approach has first been used to analyze data from the HIP randomized trial of *breast cancer screening*, which was carried out between 1963 and 1970. Parameter estimates were derived for the sensitivity of mammography and physical examination, and for the duration of preclinical screen-detectable breast cancer (Chapter II). Values between 50% and 80% for the combined sensitivity of the two tests give a good explanation of the HIP results. The mean duration of the preclinical stage is between 1.6 years for high sensitivity values and 2.7 years for low values. In comparison with previous analyses of the HIP data, our estimate for the sensitivity is lower, and the mean duration of the preclinical stage is longer. This is a consequence of the use of a more detailed model in our analysis, allowing for a more complete use of the HIP-data in testing model assumptions. The model is also in good agreement with the observed difference in breast cancer mortality between study and control group of the HIP trial.

This breast cancer screening model, which reflects the situation in the USA around 1970, has been translated to the current situation in The Netherlands, in order to evaluate the costs and effects of breast cancer screening. This extended and refined model has been checked against data from the pilot projects for breast cancer screening in Nijmegen and Utrecht (Chapter III). When the appropriate screening policy is specified, the model reproduces the detection rates and the incidence of interval cancers as observed in these projects. In comparison with the HIP model parameter values, the principal differences are the higher sensitivity of mammography, and the age-dependency of the duration of the preclinical screen-detectable stage. The sensitivity is especially high (approximately 95%) for tumours larger than 1 cm. The average duration is approximately 2 years at age 40 and increases to approximately 5 years at age 70. The model-predicted mortality reduction is in accordance with the results of the Kopparberg/Ostergötland randomised trial in Sweden.

The breast cancer screening model has subsequently been utilized in prospective evaluation of effects and costs of various screening policies in The Netherlands. The outcomes for the proposed Dutch policy of 2-yearly screening between ages 50 and 69 were favourable: an acceptable cost per life year gained, and relatively modest adverse effects that are more or less compensated by the favourable effects which occur in addition to the life-years gained. These findings have contributed considerably to the positive decision to start nationwide screening. At present, the model is being used in monitoring the results of this national programme. Subsequent cost-effectiveness analyses have been carried out and are being done for other countries and regions, including Australia, Germany, Florence, and Catalonia.

A statistical/numerical model of *cervical cancer screening* was constructed, and applied for testing hypotheses about the natural history of cervical cancer, especially about progression and regression of preclinical lesions (dysplasia and carcinoma in situ). Three models are considered and checked against data from the screening programme in British Columbia (Chapter IV). The data cover two birth cohorts and the period 1950-1969. A model without regression does not give an adequate fit of the data ($p<0.001$), and results in an implausible estimate of 33 years for the mean duration of pre-invasive lesions. A model with an equal regression rate at all ages still does not result in a good reproduction of the data. A good fit is achieved for the model that has different regression rates in lesions that develop under and over age 34. Under age 34, 84% of the new lesions will regress spontaneously, with a 95% confidence interval of 76%-92% regression. Over age 34, we estimate that 40% of the new lesions will regress. Testing more detailed age-dependency assumptions would require additional data. The average duration of dysplasia + CIS is 11.8 years, and the sensitivity of the Pap-smear is 80%. It is concluded that a considerable proportion of pre-invasive lesions in young women do not progress. The findings about progression and duration of pre-invasive lesions indicate that cervical cancer screening in young women has substantial adverse effects.

The mean total duration of the preclinical stages is estimated at about 16 years. Similar estimates had been obtained in another model-based analysis of cervical cancer and screening based on Swedish data. In an analysis of data from 10 large screening programmes, an International Agency for Research on Cancer (IARC) working group observed the build-up of incidence of invasive cervical cancer after one or more negative smears, and the results suggest a median durations of 5-10 years. An investigation has been made into the differences between the estimates for the duration of preclinical cervical cancer resulting from these two types of studies (Chapter V). Our British Columbia model has been used to predict the build-up of the incidence of invasive cancer after one and after two negative smears. The model predictions appear to correspond closely to the observed incidence trends following negative smears. The seemingly contradiction between model estimates and observed data is explained by recognizing that many of the women who have had negative smears will have further Pap smears, resulting in earlier diagnosis of invasive cervical cancers and thus an apparent faster build-up of the incidence. When the impact of further Pap smears is neglected, the data suggest that the risk of invasive cancer following one or more negative smears return close to pre-screening levels within 6-10 years. This is an overestimation of the risk of clinical invasive cancer. In case of cessation of screening it will take longer before the incidence of clinical cancer will increase. And in case of continuous screening the screen-detected cancers have a relatively favourable prognosis, thus contributing less to the serious morbidity and mortality risks associated with invasive cancer, which should be taken into account in making comparisons with the pre-screening situation.

The latter finding has been investigated in more detail. The reduction in risk of death (i.e., lethal invasive cancer) from cervical cancer after two or more negative smear test results has been calculated on basis of the long-term lethality risk of both screen-detected and clinically diagnosed cancers (Chapter VI). In women with two negative smear results estimates of protection against cervical cancer were about 50% higher when lethal invasive cancer was used as the criterion rather than all invasive cancer. This difference was due to these women being more likely to attend for further tests at which interval cancer could be detected: screen detected cancer has a better prognosis than clinically diagnosed cancer. Screening intervals could be longer than three years: screening women aged 35-64 every five years was predicted to result in a 90% reduction in mortality from cervical cancer. Because protection from mortality is higher than from disease and because of the high costs and negative side effects of frequent screening, it is concluded that screening intervals should be longer than three years.

The British Columbia model has been implemented in MISCAN and adapted to the situation in The Netherlands, and extended and refined in order to investigate the cervical cancer screening data from the last decades, and to compare a large range of different future screening policies. Transfer to MISCAN was necessary in view of the complex screening patterns which have occurred in the Dutch population and other complications such as birth cohort related trends in risk of cervical cancer. The cost-effectiveness comparisons indicated that efficient policies are characterized by a broad age-range, starting at ages 25-30 and continuing to beyond age 60. This was quite different from the prevailing policy in The Netherlands in which women were invited between age 35 and 53 with 3-years intervals. Our findings have had a major influence on the recent revision of this policy. Screening now starts at age 30 and ends at age 60, with 5 years intervals. We will be involved in continuous evaluation of the Dutch screening. Special attention will be given to evaluation of the possibilities of new screening test based on detection of Human

Papilloma Virus (HPV), that could serve as an addition or substitute of cytologic screening.

Why did the outcomes of the project "decision making in mass screening for cancer" made a real impact on actual decision making? To a large extent this may be attributed to the comprehensiveness of the evaluation, in which public health effects, experience of screening projects, diagnostic and clinical aspects, economic considerations, and epidemiological data and knowledge have each been carefully investigated and fully integrated into MISCAN cost-effectiveness models. This thesis describes some of the fundamental elements of these models.

REFERENCES

Adami HO, Malker B, Holmberg L, Persson I, Stone B. The relation between survival and age at diagnosis in breast cancer. N Engl J Med 1986;315:559-563.

Adami, HO, Gustafsson L. Cervical cancer screening (Reply). Br J Cancer 1990;62:334-335.

Albert A. Estimated cervical cancer disease state incidence and transition rates. J Natl Cancer Inst 1981;67:571-576.

Alexander FE. Estimation of sojourn time distributions and false negative rates in screening programmes which use two modalities. Stat Med 1989a;8:743-755.

Alexander FE. Statistical analysis of population screening. Med Lab Sci 1989b;46:255-267.

American Cancer Society. The cancer-related health check-up. CA 1980; 30:1-39.

Anderson GH, Boyes DA, Benedet JL, Le Riche JC, Matisic JP, Sven KC, Worth AJ, Millner A, Bennett OM. Organisation and results of the cervical cytology screening programme in British Columbia, 1955-85. Br Med J 1988;296:975-978.

Andersson I, Fagerberg G, Lundgren B, Tabar L. Breast cancer screening in Sweden. The single modality approach. Radiologe 1980;20:608-611.

Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. Br Med J 1988;297:943-948.

Baker SG, Connor RJ, Prorok PC. Recent developments in cancer screening modeling. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Cancer Screening. Cambridge University Press, Cambridge, 1991. pp 404-418.

Ballegooijen M van, Koopmanschap MA, Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Agt HMA van. Diagnostic and treatment procedures induced by cervical cancer screening. Eur J Cancer 1990;26:941-945.

Ballegooijen M van, Habbema JDF, Oortmarssen GJ van, Koopmanschap MA, Lubbe JThN, Agt HMA van. Preventive Pap-smears: striking the balance between costs, risks, and benefits. Br J Cancer 1992;65:930-933.

Ballegooijen M van, Boer R, Oortmarssen GJ van, Koopmanschap MA, Lubbe JThN, Habbema JDF. Mass screening for cervical cancer: age-ranges and intervals (in Dutch). Report MGZ.93.15. Erasmus University, Rotterdam, 1993.

Barron BA, Richart RM. A statistical model of the natural history of cervical carcinoma based on a prospective study of 557 cases. J Natl Cancer Inst 1968;41:1343-1353.

Beemsterboer PMM, Koning HJ de, Warmerdam PG, Boer R, Swart E, Dierks ML, Robra BP, Prediction of effects and costs of breast cancer screening in Germany. Int J Cancer 1994;58:623-628.

Berget A. Influence of population screening on morbidity and mortality of cancer of the uterine cervix in Maribo Amt. Dan Med Bull 1979;26:91-100.

Blumenson LE. Detection of disease with periodic screening: transient analysis and application to mammography examination. Math Biosci 1977;33:73-106.

Boer R, Warmerdam P, Koning HJ de, Oortmarssen GJ van. Extra incidence caused by mammographic screening (letter). Lancet 1994a;343;979.

Boer R, Koning HJ de, Oortmarssen GJ van, Maas PJ van der. In search for the best upper age limit for breast cancer screening. (1994b, submitted).

Boyes DA, Worth AJ, Anderson GH. Experience with cervical screening in British Columbia. Gynecol Oncol 1981;12:143-155.

Boyes DA, Morrison B, Knox EG, Draper GJ, Miller AB. A Cohort Study of Cervical Cancer Screening in British Columbia. Clin Invest Med 1982;5,1-29.

Brekelmans CTM, Collette HJA, Collette C, Fracheboud J, Waard F de. Breast cancer after a negative screen: follow-up of women participating in the DOM screening programme. Eur J Cancer 1992;28A:893-895.

Breslow L, Thomas LB, Upton AC. Final reports of the National Cancer Institute ad hoc working groups on mammography in screening for breast cancer. J Natl Cancer Inst 1979;59:467-541.

Breslow L, Thomas LB, Upton A. Predicting the benefit of screening for disease. J Appl Prob 1981;18:348-360 .

Brinton LA, Fraumeni Jr JF. Epidemiology of uterine cervical cancer. J Chron Dis 1986;39:1051-1065.

Brookmeyer R, Day NE, Moss S. Case-control studies for estimation of the natural history of preclinical disease from screening data. Stat Med 1986;5:127-138.

Brookmeyer R, Day NE. Two-stage models for the analysis of cancer screening data. Biometrics 1987;43:657-669.

Bross IDJ, Blumenson LE. Statistical testing of a deep mathematical model for human breast cancer. J Chron Dis 1968;21:493-506.

Cancer Registry of Norway. Survival of cancer patients. Cases diagnosed in Norway 1968-75. Oslo: The Norwegian Cancer Registry, 1980.

Carter R, Glasziou P, Oortmarssen GJ van, Koning HJ de, Stevenson C, Salkeld G, Boer R. Cost-effectiveness of mammographic screening in Australia. Austr J Public Health 1993;17:42-50.

Chu KC, Smart CR, Tarone RE. Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan Clinical Trial. J Natl Cancer Inst 1988;80:1125-1132.

Clarke EA, Anderson TW. Does screening by 'Pap' smears help prevent cervical cancer? Lancet 1979;II,1-4.

Clarkson DB, Wolfson DB. A displaced Poisson proces applied to screening for breast cancer in stratified samples with false negatives. Stat Neerlandica 1985;39:361-373.

Coebergh JWW, Crommelin MA, Kluck HM, Beek M van, et al. Breast cancer in south-eastern Noord-Brabant and northern Limburg: trends in earlier diagnosis in an unscreened female population. Ned Tijdschr Geneesk 1990;134:760-765. (in Dutch)

Collette C, Collette HJA, Fracheboud J, Slotboom BJ, Waard F de. Evaluation of a breast cancer screening programme - the DOM project. Eur J Cancer 1992;28A:1985-1988.

Collette HJA, Day NE, Rombach JJ, Waard F de. Evaluation of screening for breast cancer in a non-randomised study (the DOM project) by means of a case-control study. Lancet 1984; I:1224-1226.

Collette HJA, Rombach JJ, Waard F de, Collette C. An update of the DOM project for the early detection of breast cancer. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 17-27, 1988.

Collette HJA, Waard F de, Rombach JJ, Collette C, Day NE. Further evidence of benefits of a (non-randomised) breast cancer screening programme: the DOM project. J Epidemiol Comm Hlth 1992;46:382-386.

Connor RJ, Chu KC, Smart CR. Stage-shift cancer screening model. J Clin Epidemiol. 1989;42:1083-1095.

Coppleson LW, Brown B. Observations on a model of the biology of carcinoma of the cervix: a poor fit between observation and theory. Am J Obstet Gynecol 1975;122:127-136.

Day NE. Effect of cervical cancer screening in Scandinavia. Obstet Gynecol 1984;63:714-718.

Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. Biometrics, 1984;40:1-14.

Day NE. The assessment of lead time and length bias in the evaluation of screening programmes. Maturitas 1985a;7:51-58.

Day NE. Estimating the sensitivity of a screening test. J Epidemiol Comm Health 1985b;39:364-366.

Day NE, Walter SD, Tabar L, Fagerberg CJG, Collette HJA. The sensitivity and lead time of breast cancer screening: a comparison of the results of different studies. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 105-109, 1988.

Day NE. Surrogate measures in the design of breast cancer screening trials. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Cancer Screening. Cambridge University Press, Cambridge, 1991. pp 392-403.

DOM. The DOM-project for early detection of breast cancer 1983-1984. Part 3. (In Dutch). Utrecht: Preventicon, 1986.

Eddy DM. Screening for cancer: Theory, analysis and design. Englewood Cliffs NJ: Prentice-Hall, 1980.

Eddy DM. Appropriateness of cervical cancer screening. Gynecol Oncol 1981;12:S168-S187.

Eddy DM. The frequency of cervical cancer screening. Comparison of a mathematical model with empirical data. Cancer 1987;60:1117-1122.

Eddy DM, Hasselblad V, McGivney W, Hendee W. The value of mammography screening in women under age 50 years. J Am Med Ass 1988;259:1512-1519

Eddy DM. Screening for cervical cancer. Ann Intern Med 1990;113:214-226.

Eddy DM, Shwartz M. Mathematical Models in Screening. In: D. Schottenfeld and J.F. Fraumeni (eds.), Cancer Epidemiology and Prevention. pp. 1075-1090, WB Saunders, Philadelphia (1982).

Elliot PM, Tatersall MHN, Coppleson M, Russell P, Wong F, Coates AS, Solomon HJ, Bannatyne PM, Atkinson KH, Murray JC. Changing character of cervical cancer in young women. Br Med J 1989;298:288-290.

Europe Against Cancer Programme. European guidelines for quality assurance in cervical cancer screening. Eur J Cancer 1993;29A,Suppl 4:S1-S38.

Evaluation Committee. Population screening for cervical cancer in The Netherlands. Int J Epidemiol 1989;18:775-781.

Feuer EJ, Wun LM. How much of the recent rise in breast cancer incidence can be explained by increases in mammography utilization? Am J Epidemiol 1992;136:1423-36.

Fidler HK, Boyes DA, Worth AJ. Cervical cancer detection in British Columbia. J Obstet Gynaec Brit Cwlth 1968;75:392-404.

Fink R, Shapiro S, Lewison J. The reluctant participant in a breast cancer screening program. Public Health Report 1968;83:479-490.

Fink R, Shapiro S, Roesner R. Impact of efforts to increase participation in repetitive screenings for early breast cancer detection. Amer J Public Health 1972;62:328-336.

Flehinger BJ, Kimmel M. The natural history of lung cancer in a periodically screened population. Biometrics 1987;43:127-144.

Flehinger BJ, Kimmel M, Polyak T, Melamed MR. Screening for lung cancer. The Mayo Lung Project revisited. Cancer. 1993;72:1573-1580.

Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the international workshop on screening for breast cancer. J Natl Cancer Inst 1993;85:1644-1656.

Forrest P. Breast cancer screening. Report to the Health Ministers of England, Wales, Scotland & Northern Ireland. London: Dept. of Health and Social Security, 1986.

Friedman DR, Dubin N. Case-control evaluation of breast cancer screening efficacy. Am J Epidemiol 1991;133:974-984.

Friedman GD, Collen MF, Fireman BH. Multiphasic health checkup evaluation: a 16-year follow-up. J Chron Dis 1986; 39:453-463.

Gad A, Thomas BA, Moskowitz M. Screening for breast cancer in Europe: achievements, problems and future. Recent Results in Cancer Research 1984;90:179-184.

Gezondheidsraad. Early detection of breast cancer (In Dutch). 's-Gravenhage: Gezondheidsraad, 1987.

Glasziou PP. Meta-analysis adjusting for compliance: the example of screening for breast cancer. J Clin Epidemiol 1992;45:1251-1256.

Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. Biometrics 1978;34:77-86.

Goldberg JD, Wittes JT. The evaluation of medical screening procedures. The American Statistician 1981;35:4-10.

Graaf Y van der. Screening for cervical cancer. The Nijmegen project (Ph.D. Thesis). Nijmegen: Katholieke Universiteit, 1987.

Graaf Y van der, Zielhuis GA, Peer PGM, Vooys GP. The effect of cervical cancer screening; a population based case-control study. J Clin Epidemiol 1988;41:21-26.

Graaf Y van der, Kallewaard M, Hof-Grootenboer AE van 't, Otto LP, Sie-Go DMDS, Woudt JMC, Vooys GP. Incidentie van invasief plaveiselcelcarcinoom bij vrouwen van 55-68 jaar die eerder in het bevolkingsonderzoek werden gescreend. Rijksuniversiteit Utrecht. Utrecht, 1991.

Gustafsson L. The natural history of cancer of the cervix uteri. A simulation study based on Swedish statistics for 1958-1981. Institute of Technology, Uppsala University. UPTEC 8607R, Uppsala, 1986.

Gustafsson L, Adami HO. Natural history of cervical neoplasia: consistent results obtained by an identification technique. Br J Cancer 1989;60:132-141.

Gustafsson L, Adami HO. Cytologic screening for cancer of the uterine cervix in Sweden evaluated by identification and simulation. Br J Cancer 1990;61:903-908.

Gustafsson L, Adami HO. Optimization of cervical cancer screening. Cancer Causes and Control 1992;3:125-136.

Habbema JDF, Jong GA de, Maas PJ van der, Oortmarssen GJ van. Besluitvorming over bevolkingsonderzoek. T soc Geneesk 1978;56:33-43.

Habbema JDF, Oortmarssen GJ van, Maas PJ van der. Mass Screening for Cancer: The Interpretation of Findings and the Prediction of Effects on Morbidity and Mortality. Clin Lab Med 1982a;2:627-638.

Habbema JDF, Oortmarssen GJ van. Performance characteristics of screening tests. Clin Lab Med 1982b;2:639-656.

Habbema JDF, Lubbe JTN, Maas PJ van der, Oortmarssen GJ van. Het MISCAN model: een nieuwe methode voor evaluatie van bevolkingsonderzoek. Deel I: bevolkingsonderzoek op baarmoederhalskanker: analyse van de British Columbia Cohort Study. Technical Report, Erasmus University Rotterdam, 1982c.

Habbema JDF, Oortmarssen GJ van, Putten DJ van. An analysis of survival differences between clinically and screen-detected cancer patients. Statistics in Medicine 1983;2:279-285.

Habbema JDF, Oortmarssen GJ van, Lubbe JThN, Maas PJ van der. The MISCAN simulation program for the evaluation of screening for disease. Comput Meth Progr Biomed 1984;20:79-83.

Habbema JDF, Oortmarssen GJ van, Lubbe JTN, Maas PJ van der. Model building on the basis of Dutch cervical cancer screening data. Maturitas 1985;7:11-20.

Habbema JDF, Oortmarssen GJ van, Putten DF van, et al. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York. J Natl Cancer Inst 1986;77:317-320.

Habbema JDF, Lubbe JThN, Oortmarssen GJ van, Maas PJ van der. A simulation approach to cost-effectiveness and cost-benefit calculations of screening for the early detection of disease. Eur J Oper Res 1987;29:159-166.

Habbema JDF, Lubbe JThN, Agt HME van, Ballegooijen M van, Koopmanschap MA, Oortmarssen GJ van. The costs and effects of mass screening for cervical cancer. (In Dutch). Rotterdam: Erasmus University, Dept. of Public Health and Social Medicine, 1988.

Habbema JDF, Oortmarssen GJ van. To screen or not to screen. How do we decide on which cancer screening activities to embark upon? Eur J Cancer 1994;30A;884-886.

Haes JHCJM, Koning HJ de, Oortmarssen GJ van, et al. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer 1991;49:538-544.

Hakama M, Miller AB, Day NE, eds. Screening for Cancer of the Uterine Cervix. Lyon: International Agency for Research on Cancer, 1986a.

Hakama M, Miller AB. Introduction: Report on the Workshop on Screening for Cancer of the Cervix. In: M Hakama, AB Miller & NE Day, eds. Screening for Cancer of the Uterine Cervix. Lyon: International Agency for Research on Cancer, 1986b:147-148.

Hakama M. Design of the Finnish breast cancer screening study. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 59-62, 1988.

Hakama M, Elovaino L, Kajantie R, Louhivuori K. Breast cancer screening as public health policy in Finland. Br J Cancer 1991;64:962-964.

Hakulinen T, Pukkala E, Hakama M, Lehtonen M, Saxen E, Teppo L. Survival of cancer patients in Finland 1953-1974. Ann Clin Res 1983;13,Suppl 31:50-52.

HIP Study. Evaluation of periodic breast cancer screening with mammography and clinical examination. Progress Report. Health Insurance Plan of Greater New York, 1981.

Holford TR, Zhang Z, McKay LA. Estimating age, period and cohort effects using the multistage model for cancer. Stat Med 1994;13:23-41.

Holland WW, Stewart S. Screening in Health Care. Nuffield Provincial Hospitals Trust. London 1990.

IARC working group on evaluation of cervical cancer screening programmes. Screening for squamous cervical cancer: duration of low risk after negative results of cervical cytology and its implication for screening policies. Br Med J 1986;293:659-664.

Ineveld BM van, Oortmarssen GJ van, Koning HJ de, Boer R, Maas PJ van der. How cost-effective is breast cancer screening in different EEC countries? Eur J Cancer 1993;29A:1663-1668.

Kinlen LJ, Spriggs AI. Women with positive cervical smears but without surgical intervention. Lancet 1978:463-465.

Kleijnen JPC, Groenendaal W van. Simulation - A Statistical Perspective. Wiley, Chichester, 1992.

Knox EG. A simulation system for screening procedures. In: McLachlan G, ed. The future and present indicatives. London: Oxford University Press, 1973:19-55.

Knox EG. Simulation studies of breast cancer screening programmes. In: G. McLachlan (ed.), Probes for Health, pp.15-44, Nuffield Provincial Hospitals Trust, London (1975)

Knox EG. Biological simulation and health care planning. Invited paper, 1975 ISI symposium, pp 1-12.

Knox EG, Woodman CBJ. Effectiveness of a cancer control programme. Cancer Surveys 1988;7:379-401.

Knox EG. Evaluation of a proposed breast cancer screening regimen. Br Med J 1988;297:650-654.

Koning HJ de, Oortmarssen GJ van, Ineveld BM van, Maas PJ van der (1990a). Breast cancer screening: its impact on clinical medicine. Br J Cancer 61:292-297.

Koning HJ de, Ineveld BM van, Oortmarssen GJ van, et al. (1990b) The costs and effects of mass screening for breast cancer. Final Report (in Dutch). Instituut Maatschappelijke Gezondheidszorg, Rotterdam.

Koning HJ de, Ineveld BM van, Oortmarssen GJ van, et al. Breast cancer screening and cost-effectiveness: policy alternatives, quality of life considerations, and the possible impact of uncertain factors. Int J Cancer 1991;49:531-537.

Koning HJ de. The effects and costs of breast cancer screening. Ph.D. Thesis, Erasmus University Rotterdam, 1993.

Koopmanschap MA, Lubbe JThN, Oortmarssen GJ van, Agt HME van, Ballegooijen M van, Habbema JDF. Economic aspects of cervical cancer screening. Soc Sci Med 1990a;30:1081-1087

Koopmanschap MA, Oortmarssen GJ van, Agt HMA van, Ballegooijen M van, Habbema JDF, Lubbe JThN. Cervical cancer screening: attendance and cost-effectiveness. Int J Cancer 1990b;45:410-415.

Koopmanschap MA. Complementary analyses in economic evaluation of health care. Ph.D. Thesis, Erasmus University Rotterdam, 1994.

Kottmeier HL. Evolution et traitement des epitheliomas. Rev Fr Gynecol Obstet 1961;56:821-826.

Kusama S, Spratt JS, Donegan WL, Watson FR, Cunningham C. The gross rate of growth of human mammary carcinoma. Cancer 1972;30:594-599.

Läärä E, Day NE, Hakama M. Trends in mortality from cervical cancer in the Nordic countries: association with organized screening programmes. Lancet 1987;I:1247-1249.

Editorial. Breast cancer screening in women under 50. Lancet 1991;337:1575-1576.

Law AM, Kelton WD. Simulation Modeling and Analysis. McGraw-Hill, New York. 1982.

LETB (Landelijk Evaluatie Team voor bevolkingsonderzoek naar Borstkanker). Landelijke Evaluatie van bevolkingsonderzoek op borstkanker I. Erasmus Universiteit, Rotterdam, 1992.

LETB (Landelijk Evaluatie Team voor bevolkingsonderzoek naar Borstkanker). Landelijke Evaluatie van bevolkingsonderzoek op borstkanker II. Erasmus Universiteit, Rotterdam, 1993.

LETB (Landelijk Evaluatie Team voor bevolkingsonderzoek naar Borstkanker). Landelijke Evaluatie van bevolkingsonderzoek op borstkanker III. Erasmus Universiteit, Rotterdam, 1994.

Louis TA, Albert A, Heghinian S. Screening for early detection of cancer - III. Estimation of disease natural history. Math Biosci 1978;41:49-54.

Lynge E, Poll P. Incidence of cervical cancer following negative smear. A cohort study from Maribo county, Denmark. Am J Epidemiol 1986;124:345-52.

Maas PJ van der, Koning HJ de, Ineveld BM van, Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Geerts AT, Collette HJA. Verbeek ALM, Hendriks JHCL, Rombach JJ. The cost-effectiveness of breast cancer screening. Int J Cancer 1989;43:1055-1060.

Macgregor JE, Moss SE, Parkin DM, Day NE. A case-control study of cervical cancer screening in North-East Scotland. Br Med J 1985;290:1543-1546.

Magnus K, Langmark F, Andersen A. Mass screening for cervical cancer in Ostfold county of Norway 1959-77. Int J Cancer 1987;39:311-316.

Miller AB, Lindsay J, Hill GB. Mortality from cancer of the uterus in Canada and its relationship to screening for cancer of the cervix. Int J Cancer 1976;17:602-612.

Miller AB. Screening for breast cancer. Breast Cancer Res Treatm 1983; 3:143-156.

Miller AB. Evaluation of the impact of screening for cancer of the cervix. In: M Hakama, AB Miller & NE Day, eds. Screening for Cancer of the Uterine Cervix. Lyon: Int Agency Res Cancer, 1986:149-160.

Miller AB, Howe GR, Wall C. The national study of breast cancer screening: protocol for a Canadian randomized controlled trial of screening for breast cancer in women. Clin Invest Med 1980;4:227-258.

Moens GFG, Oortmarssen GJ van, Honggokoesoemo S, Voorde H van de. Birth cohort analysis of suicide mortality in Belgium 1954-81 by a graphic and a quantitative method. Acta Psychiatr Scand 1987;76:450-455.

Morrison AS. Case definition in case-control studies of the efficacy of screening. Am J Epidemiol 1982;115:6-8.

Morrison AS. Screening in Chronic Disease, second edition. Oxford University Press, New York / Oxford. 1992.

Moss SM. Case-control studies of screening. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Cancer Screening. Cambridge University Press, Cambridge, 1991. pp 419-428.

NCI Working guidelines for early cancer detection 1987.

NCHS. Vital Statistics of the United States. US Dept. of HEW, Rockville, (1980).

NCHS. New York State Life Tables: 1969-71. US Dept. of HEW, Rockville, (1975).

Netherlands Central Bureau of Statistics. Life tables for the Netherlands, 1982-83. Monthly Bulletin of Population Statistics 1985;85/4:44-56.

Netherlands Central Bureau of Statistics. Cancer morbidity and mortality 1981-1982. Monthly Bulletin of Health Statistics 1985;85/3:5-25.

Nyström L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Rydén S, Andersson I, Bjurstam N, Fagerberg G, Frisell J, Tabár L, Larsson L-G. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;341:973-978.

Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Jong GA de, Maas PJ van der. Predicting the effects of mass screening for disease - a simulation approach. Eur J Oper Res 1982;6:399-409.

Oortmarssen GJ van, Habbema JDF. Cervical cancer screening data from two cohorts in British Columbia. In: M Hakama, AB Miller & NE Day, eds. Screening for Cancer of the Uterine Cervix. Lyon: Int Agency Res Cancer, 1986:47-60.

Oortmarssen GJ van, Habbema JDF (1990). Cervical cancer screening (Letter to the editor). Br J Cancer 1990;62:333.

Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Maas PJ van der. A model-based analysis of the HIP-project for breast cancer screening. Int J Cancer 1990;46:207-213.

Oortmarssen GJ van, Habbema JDF, Maas PJ van der, Koning HJ de, Collette HJA, Verbeek ALM, Geerts AT, Lubbe JThN. A model for breast cancer screening. Cancer 1990;66:1601-1612.

Oortmarssen GJ van, Habbema JDF. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. Br J Cancer 1991;64:559-564.

Oortmarssen GJ van, Habbema JDF, Ballegooijen M van. Predicting mortality from cervical cancer after negative smear test results. Br Med J 1992;305:449-451.

Oortmarssen GJ van, Ermoliev Y. On Stochastic optimization of disease control strategies under risk and uncertanty. Working Paper. International Institute for Applied Systems Analysis, Laxenburg, Austria (1994).

Osmond C, Gardner MJ. Age, Period and Cohort models applied to cancer mortality rates. Stat Med 1982;1:245-259.

Paci E, Duffy SW. Modelling the analysis of breast cancer screening programmes: sensitivity, lead time and predictive value in the Florence District Programme (1975-1986). Int J Epidemiol 1991;20:852-858.

Paci E, Boer R, Zappa M, Koning HJ de, Oortmarssen GJ van, Crocetti E, Giorgi D, Roselli del Turco M, Habbema JDF. Model based predictions of the impact on mortality reduction of a mammographic screening program in the city of Florence. (submitted)

Palli D, Roselli Del Turco M, Buiatti E, Carli S, Ciatto S, Toscani I, Maltoni G. A case-control study of the efficacy of a non-randomized breast cancer screening programme in Florence (Italy). Int J Cancer 1986;38:501-504.

Pamilo M, Dean PB, Räsänen O. Mammography screening for breast cancer: four years of nationwide screening in Finland. Eur Radiol 1993;3:44-45.

Parkin DM. A computer simulation model for the practical planning of cervical cancer screening programmes. Br J Cancer 1985;51:551-568.

Parkin DM, Moss SM. An evaluation of screening policies for cervical cancer in England and Wales using a simulation model. J Epidemiol Comm Hlth 1986;40:143-153.

Peeters PMH, Verbeek ALM, Hendriks JHCL, Bon MJH van. Screening for breast cancer in Nijmegen: report of 6 screening rounds, 1975-1986. Int J Cancer 1989;43:226-230.

Peer PGM, Dijck JAAM, Hendriks JHCL, Holland R, Verbeek ALM. Age-dependent growth rate of primary breast cancer. Cancer 1993;71:3547-3551.

Peer PGM, Holland R, Hendriks JHCL, Mravunac M, Verbeek ALM. Age-specific effectiveness of the Nijmegen population-based breast cancer screening program: Assessment of early indicators of screening effectiveness. J Natl Cancer Inst 1994;86:436-441.

Pelikan S, Moskowitz M. Effects of lead time, length bias, and false-negative assurance on screening for breast cancer. Cancer 1993;71:1998-2005.

Plaisier AP, Alley ES, Boatin BA, Oortmarssen GJ van, Remme J, Vlas SJ de, Habbema JDF. Irreversible effects of Ivermectin on adult parasites in onchocerciasis patients in the Onchocerciasis Control Programme in West Africa. Submitted for publication.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. (1988). Numerical recipes in C. The art of scientific computing. Cambridge University Press: Cambridge.

Prorok PC. The theory of periodic screening. I: Lead time and proportion detected. Adv Appl Prob 1976a;8:127-143.

Prorok PC. The theory of periodic screening. II: Doubly bounded recurrence times and mean lead time and detection probability estimation. Adv Appl Prob 1976b;8:460-476.

Prorok PC, Hankey BF, Bundy BN. Concepts and problems in the evaluation of screening programs. J Chron Dis 1981;34:159-171.

Prorok PC. Mathematical models and natural history in vervical vancer screening. In: M Hakama, AB Miller & NE Day, eds. Screening for Cancer of the Uterine Cervix. Lyon: Int Agency Res Cancer 1986:185-196.

Prorok PC. Mathematical models of breast cancer screening. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 195-104, 1988.

Putten DJ van, Habbema JDF, Oortmarssen GJ van, Maas PJ van der (1981). Preliminaries for evaluation of breast cancer screening. Technical Report, Dept of Public Health and Social Medicine, Erasmus University, Rotterdam.

Ripley BD. Stochastic simulation. Wiley, New York, 1987.

Roberts MM, Alexander FE, Anderson TJ, Chetty U, Donnan PT, Forrest P, Hepburn W, Huggins A, Kirkpatrick AE, Lamb J, Muir BB, Prescott RJ. Edinburg trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241-246.

Rutqvist LE. On the utility of the lognormal model for analysis of breast cancer survival in Sweden, 1961-1973. Br J Cancer 1985;52:875-883.

Rutqvist LE, Miller AB, Andersson I, Hakama M, Hakulinen T, Sigfússon BF, Tabár L. Reduced breast cancer mortality with mammography screening - an assessment of currently available data. Int J Cancer Supplement 1990;5:76-84.

Sasco AJ, Day NE, Walter SD. Case-control studies for the evaluation of screening. J Chron Dis 1986;39:399-405.

Screening Brief. Breast Cancer. Journal of Medical Screening 1994;1:73.

Shapiro S. Evidence on screening for breast cancer from a randomized trial. Cancer 1977; 39:2772-2782.

Shapiro S, Goldberg JD, Hutchinson GB. Lead time in breast cancer detection and implications for periodicity. Amer. J. Epidemiology 1974;100:357-366.

Shapiro S, Venet W, Strax P, Venet L. Current results of the breast cancer screening randomized trial: the Health Insurance Plan (HIP) of Greater New York. In: N.E. Day and A.B. Miller (eds.), Screening for Breast Cancer, pp. 3-16, H. Huber Publishers, Toronto (1988).

Shapiro S, Strax P, Venet L, Venet W. Changes in 5-years breast cancer mortality in a breast cancer screening program. Seventh National Cancer Conference Proceedings, Am Cancer Soc 1974:663-678.

Shapiro S, Venet L. Evaluation of periodic breast cancer screening with mammography. JAMA 1966;195:111-118.

Shapiro S. Statistical evidence for mass screening for breast cancer and some remaining issues. Cancer Detection and Prevention 1976;1:347-363.

Shapiro S, Strax P, Venet L. Periodic breast cancer screening. In: Sharp CL, Keen H, eds. Presymptomatic Detection and Early Diagnosis. London: Pitman Med Publ 1968:203-235.

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

Sherlaw-Johnson C, Gallivan S, Jenkins D, Jones MH. Cytological screening and management of abnormalities in prevention of cervical cancer: an overview with stochastic modelling. J Clin Pathol 1994;47:430-435.

Shwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. Cancer 1978a;41:1550-1564.

Shwartz M. A mathematical model used to analyze breast cancer screening strategies. Operations Res 1978b;26:937-955.

Shwartz M. Estimates of lead time and length bias in a breast cancer screening program. Cancer 1980;46:844-851.

Shwartz M. Validation and use of a mathematical model to estimate the benefits of screening younger women for breast cancer. Cancer Detect Prev 1981;4:595-601.

Shwartz M. Validation of a model of breast cancer screening. An outlier observation suggests the value of breast self-examination. Med Decis Making 1992;12:222-228.

Silcocks PBS, Moss SM. Rapidly progressive cervical cancer: is it a real problem? Br J Obset Gynecol 1988;95:1111-1116.

Stevenson CE, Glasziou P, Carter R, Fett MJ, Oortmarssen GJ van. Using computer modelling to estimate person years of life saved by mammography screening in Australia. In: C Selby Smith (ed) Economics and Health:1990. Proceedings of the twelfth Australian conference on health economics. Monash University, Clayton, Victoria, 1991, pp 295-307.

Tabar L, Gad A, Eklund G, et al. Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1985;I:829-832.

Tabar L, Duffy SW, Krusemo UB. Detection method, tumour size and node metastases in breast cancers detected during a trial of breast cancer screening. Eur J Cancer 1987;23:959-962.

Tabar L, Fagerberg D, Day NE, Holmberg L. 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 1987;55:547-551.

Tabar L, Fagerberg CJG, Day NE. The results of periodic one-view mammography screening in a randomized, controlled trial in Sweden. Part 2: Evaluation of the results. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 39-44, 1988.

Tabar L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Comm Health 1989;43:107-114.

TNCS. Third National Cancer Survey: Incidence Data. NCI Monograph no. 41. National Cancer Institute, Bethesda, MD.

UK trial of early detection of breast cancer group. First results on mortality reduction on the UK trial of early detection of breast cancer. Lancet 1988;II:411-416.

Verbeek ALM, Straatman H, Hendriks JHCL. Sensitivity of mammography in Nijmegen women under age 50: some trials with the Eddy model. In: Day NE and Miller AB (eds.), Screening for Breast Cancer. Toronto: H. Huber Publishers, pp. 29-32, 1988.

Verbeek ALM, Hendriks JHCL, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of cancer mortality through mass screening with modern mammography: first results of the Nijmegen project 1987-1981. Lancet 1984; I:1222-1224.

Verhoeven L. Bevolkingsonderzoek op baarmoederhalskanker. Het bepalen van de beste leeftijden waarop vrouwen zouden moeten worden uitgenodigd voor deelname aan het bevolkingsonderzoek op baarmoederhalskanker. Vakgroep Mathematische besliskunde, Economische Faculteit, Rotterdam, 1994.

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

Walter SD, Day NE. Estimation of the duration of the pre-clinical state using screening data. Am J Epidemiol 1983;118:865-886.

Walter SD, Stitt LW. Evaluating the survival of cancer cases detected by screening. Stat Med 1987;6:885-900.

Walton RJ. Cervical Cancer Screening Programs: Summary of the 1982 Canadian Task Force Report. CMA J 1982;127:581-589.

Weiss NS. Control definition in case-control studies of the efficacy of screening and diagnostic testing. Am J Epidemiol 1983;118:457-460.

Weis NS. Application of the case-control method in the Evaluation of Screening. Epidemiologic Reviews 1994;16:102-108.

Yu SZ, Miller AB, Sherman GJ. Optimising the age, number of tests, and test interval for cervical cancer screening in Canada. J Epidemiol Comm Hlth 1982;36:1-10.

Zelen M, Feinleib M. On the theory of screening for chronic diseases. Biometrika 1969;56:601-613.

Zelen M. Optimal scheduling of examinations for the early detection of disease. Biometrika 1993;80:279-93.

LIST OF PUBLICATIONS ON DISEASE MODELLING.

Mass screening for cancer

Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Jong GA de, Maas PJ van der. Predicting the effects of mass screening for disease - a simulation approach. Eur J Oper Res 1982;6:399-409.

Habbema JDF, Oortmarssen GJ van, van Putten DJ. An analysis of survival differences between clinically and screen-detected cancer patients. Statistics in Medicine 1983;2:279-285.

Habbema JDF, Oortmarssen GJ van, Lubbe JThN, Maas PJ van der. The MISCAN simulation program for the evaluation of screening for disease. Comp Meth Progr Biomed 1984;20:79-93.

Habbema JDF, Oortmarssen GJ van, Lubbe JTN, Maas PJ van der. Model building on the basis of Dutch cervical cancer screening data. Maturitas 1985;7:11-20.

Habbema JDF, Oortmarssen GJ van, Putten DF van, et al. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York. J Natl Cancer Inst 1986;77:317-320.

Habbema JDF, Lubbe JThN, Oortmarssen GJ van, Maas PJ van der. A simulation approach to cost-effectiveness and cost-benefit calculations of screening for the early detection of disease. Eur J Oper Res 1987;29:159-166.

Maas PJ van der, Koning HJ de, Ineveld BM van, Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Geerts AT, Collette HJA. Verbeek ALM, Hendriks JHCL, Rombach JJ. The cost-effectiveness of breast cancer screening. Int J Cancer 1989;43:1055-1060.

Ballegooijen M, Koopmanschap MA, Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Agt HMA van. Diagnostic and treatment procedures induced by cervical cancer screening. Eur J Cancer 1990;26:941-945.

Koning HJ de, Oortmarssen GJ van, Ineveld BM van, Maas PJ van der. Breast cancer screening: its impact on clinical medicine. Br J Cancer 1990;61:292-297.

Koopmanschap MA, Oortmarssen GJ van, Agt HMA van, Ballegooijen M van, Habbema JDF, Lubbe JThN. Cervical cancer screening: attendance and cost-effectiveness. Int J Cancer 1990;45:410-415.

Koopmanschap MA, Lubbe JThN, Oortmarssen GJ van, Agt HME van, Ballegooijen M van, Habbema JDF. Economic aspects of cervical cancer screening. Soc Sci Med 1990;30:1081-1087

Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Maas PJ van der. A model-based analysis of the HIP-project for breast cancer screening. Int J Cancer 1990;46:207-213.

Oortmarssen GJ van, Habbema JDF, Maas PJ van der, Koning HJ de, Collette HJA, Verbeek ALM, Geerts AT, Lubbe JThN. A model for breast cancer screening. Cancer 1990;66:1601-1612.

Oortmarssen GJ van, Habbema JDF. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. Br J Cancer 1991;64:559-565.

Koning HJ de, Ineveld BM van, Oortmarssen GJ van, Haes JHCJM de, Collette HJA, Hendriks JHCL, Maas PJ van der. Breast cancer screening and cost-effectiveness:

policy alternatives, quality of life considerations, and the possible impact of uncertain factors. Int J Cancer **1991**;49:531-537.

Haes JHCJM de, Koning HJ de, Oortmarssen GJ van, Bruyn AE de, Maas PJ van der. The impact of a breast cancer screening programme on quality-adjusted life-years. Int J Cancer **1991**;49:538-544.

Koning HJ de, Ineveld BM van, Haes JHCJM de, Oortmarssen GJ van, Klijn JGM, Maas PJ van der. Advanced breast cancer and its prevention by screening. Br J Cancer **1992**;65:950-955.

Ballegooijen M van, Habbema JDF, Oortmarssen GJ van, Koopmanschap MA, Lubbe JThN, Agt HMA van. Preventive Pap-smears: striking the balance between costs, risks, and benefits. Br J Cancer **1992**;65:930-933.

Ballegooijen M van, Koopmanschap MA, Subandono Tjokrowardojo AJ, Oortmarssen GJ van. Care and costs for advanced cervical cancer. Eur J Cancer **1992**:28A:1703-1708.

Oortmarssen GJ van, Habbema JDF, Ballegooijen M van. Predicting mortality from cervical cancer after negative smear test results. Br Med J **1992**;305:449-451.

Oortmarssen GJ van, Habbema JDF, Ballegooijen M van. Predicting mortality from cervical cancer (letter). Br Med J **1992**;305:833.

Carter R, Glasziou P, Oortmarssen GJ van, Koning HJ de, Stevenson C, Salkeld G, Boer R. Cost-effectiveness of mammographic screening in Australia. Australian Journal of Public Health **1993**;17:42-50.

Ineveld BM van, Oortmarssen GJ van, Koning HJ de, Boer R, Maas PJ van der. How cost-effective is breast cancer screening in different EEC countries? Eur J Cancer **1993**;29A:1663-1668.

Boer R, Warmerdam P, Koning HJ de, Oortmarssen GJ van. Extra incidence caused by mammographic screening (letter). Lancet **1994**;343;979.

Habbema JDF, Oortmarssen GJ van. To screen or not to screen. How do we decide on which cancer screening activities to embark upon? Eur J Cancer **1994**;30A;884-886.

Oortmarssen GJ van, Habbema JDF. The duration of pre-clinical cervical cancer and the reduction in incidence of invasive cancer following negative Pap-smears. Int J Epidemiol *(accepted for publication)*

Various

Moens GFG, Oortmarssen GJ van, Honggokoesoemo S, Voorde H van de. Birth cohort analysis of suicide mortality in Belgium 1954-81 by a graphic and a quantitative method. Acta Psychiatr Scand **1987**;76:450-455.

Beeck EF van, Mackenbach JP, Oortmarssen GJ van, Barendregt J, Habbema JDF, Maas PJ van der. Scenarios for the future development of accident mortality in The Netherlands. Health Policy **1989**;11:1-17.

Deeg DJH, Oortmarssen GJ van, Habbema JDF, Maas PJ van der. A measure of survival time for long-term follow-up studies of the elderly. J Clin Epidemiol **1989**;42:541-549.

Benbassat J, Zajicek G, Oortmarssen GJ van, Ben-Dov- I, Eckmann MH. Inaccuracies in estimates of life-expectancies of patients with bronchial cancer in clinical decision making. Med Decis Making 1993;13:237-244.

Bonneux L, Oortmarssen GJ van, Barendregt JJ. Increase in cancer incidence in younger birth cohorts (letter). Lancet 1993;341:1409.

Control of river blindness

Plaisier AP, Oortmarssen GJ van, Habbema JDF, Remme J, Alley ES. ONCHOSIM: a model and computer simulation program for the transmission and control of onchocerciasis. Comp Methods and Programs in Biomed 1990; 31:43-56

Habbema JDF, Plaisier AP, Oortmarssen GJ van, Remme J. Prospective evaluation of onchocerciasis control strategies. Acta Leidensia 1990;59:387-398.

Remme J, Sole G de, Oortmarssen GJ van. The predicted and observed decline in onchocerciasis infection during 14 years of successful simulium control in West-Africa with reference to the reproductive lifespan of onchocerca volvulus. Bull. WHO 1990;68:331-339.

Plaisier AP, Oortmarssen GJ van, Remme JHF, Habbema JDF. The reproductive lifespan of *Onchocerca volvulus* in West African Savanna. Acta Tropica 1991;48:271-284.

Plaisier AP, Oortmarssen GJ van, Remme JHF, Alley ES, Habbema JDF. The risk and dynamics of onchocerciasis recrudescence after cessation of vector control. Bull. WHO 1991;69:169-178.

Habbema JDF, Alley ES, Plaisier AP, Oortmarssen GJ van, Remme JHF. Epidemiological modelling for onchocerciasis control. Parasitology Today 1992;8:99-103.

Control of Schistosomiasis

Vlas SJ de, Gryseels B, Oortmarssen GJ van, Polderman AM, Habbema JDF. A model for variations in single and repeated egg counts in *Schistosoma mansoni* infections. Parasitology 1992;104:451-460.

Vlas SJ de, Oortmarssen GJ van, Gryseels B. Validation of a model for variations in *Schistosoma mansoni* egg counts. Transactions of the Royal Society of Tropical Medicine and Hygiene, 1992;86:645.

Vlas SJ de, Nagelkerke NJD, Habbema JDF, Oortmarssen GJ van. Statistical Models for estimating prevalence and incidence of parasitic diseases. Statistical Methods in Medical Research, 1993;2:3-21.

Vlas SJ de, Gryseels B, Oortmarssen GJ van, Polderman AM, Habbema JDF. A Pocket Chart to Estimate True *Schistosoma mansoni* prevalences. Parasitology Today 1993;9:305-307.

Leprosy control

Habbema JDF, Jozefzoon E, Oortmarssen GJ van. Towards the use of decision sciences in leprosy control. Leprosy Review 1992;63:Supplement,48s-52s.

ACKNOWLEDGMENTS

I wish to thank the persons who provided us the screening data sets and assisted in analysing and interpreting the data:
The British Columbia cervical cancer screening data (drs AB Miller and DA Boyes)
The statistical tape of the HIP breast cancer screening trial (mrs Wanda Venet)
The DOM project for breast cancer screening (prof Bertien Collette)
The Nijmegen study of breast cancer screening (prof Andre Verbeek)

Model-based evaluation of cancer screening was financed by the Ministry of Health, Welfare and Cultural Affairs, the Dutch Cancer Foundation (Koningin Wilhelmina Fonds), The Prevention Fund, and the Public Health Insurance Council.

Chapters II-VI have been reused with permission:

Oortmarssen GJ van, Habbema JDF, Lubbe JThN, Maas PJ van der. A model-based analysis of the HIP-project for breast cancer screening. Int J Cancer **1990**;46:207-213.

Oortmarssen GJ van, Habbema JDF, Maas PJ van der, Koning HJ de, Collette HJA, Verbeek ALM, Geerts AT, Lubbe JThN. A model for breast cancer screening. Cancer **1990**;66:1601-1612.

Oortmarssen GJ van, Habbema JDF. Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer. Br J Cancer **1991**;64:559-565.

Oortmarssen GJ van, Habbema JDF, van Ballegooijen M. Predicting mortality from cervical cancer after negative smear test results. Br Med J **1992**;305:449-451.

Oortmarssen GJ van, Habbema JDF. The duration of pre-clinical cervical cancer and the reduction in incidence of invasive cancer following negative Pap-smears. *Int J Epidemiol.*

DANKWOORD

Dit proefschrift beschrijft een onderdeel van een langlopend multidisciplinair project waaraan in de loop der tijd velen een belangrijke bijdrage hebben geleverd.

In de eerste plaats ben ik dank verschuldigd aan Dik Habbema, Paul van der Maas en Gaspard de Jong die rond 1977 besloten om het project "Besluitvorming over Bevolkingsonderzoek" te starten, en in mij een geschikte persoon zagen om dit onderzoek uit te voeren. De visie en daadkracht van Paul en Dik waren en zijn nog steeds cruciaal voor het welslagen van dit project, en daarmee ook voor het totstand komen van dit proefschrift. Dank zij hun voortdurende aansporingen is dit boek er dan toch nog gekomen.

Gedurende vele jaren vormden Koos Lubbe en ik de kern van het project. Wat ik vooral in Koos heb gewaardeerd is dat hij naast het MISCAN programmeerwerk ook goed oog had voor de continuiteit van het project, bijvoorbeeld bij het veiligstellen van toegangstijd en rekentijd op rekencentra, en bij het goed doorstaan van veranderingen van hardware. Graag wil ik ook Dick van Putten bedanken, die als eerste arts-projectmedewerker een zeer aanzienlijke bijdrage geleverd aan de (MISCAN) analyse van het HIP project (hoofdstuk II).

Een grote stimulans voor het schrijven van publikaties, en daarmee voor het tot stand komen van dit proefschrift, vormde de zeer prettige en vruchtbare samenwerking met de medewerkers van de kosten-effectiviteitsprojecten. In dit verband gaat mijn dank uit naar Marjolein van Ballegooijen, Harry de Koning, Martin van Ineveld, Marc Koopmanschap, Hanneke de Haes, Heleen van Agt, Ada Geerts, Arry de Bruijn, en niet in de laatste plaats Rob Boer die langzamerhand een groot deel van mijn modelbouw-aktiviteiten rond bevolkingsonderzoek heeft overgenomen. Ook aan de constructieve contacten met de Nijmeegse en Utrechtse partners in MACEA verband, zonder wie hoofdstuk III niet gemaakt had kunnen worden, bewaar ik goede herinneringen.

Buiten de SCREEN groep heb ik veel gehad aan de methodologische discussies met Anton Plaisier en Sake de Vlas rond data analyse en microsimulatie, en met Luk Bonneux en Jan Barendregt over de (on)mogelijkheden van modellering van ziekteprocessen.

Veel waardering heb ik voor de mensen van het MGZ secretariaat, in het bijzonder Saskia Drent, die met engelengeduld van mijn gecompliceerde en veelvormige tekstbestanden een uniform geheel heeft weten te maken.

Boven alles gaat mijn dank uit naar mijn ouders die het mogelijk hebben gemaakt om "door te leren", en aan Fransien, Dagmar en Hainuwele die het vaak zonder mijn aanwezigheid hebben moeten stellen, was het niet vanwege verblijf in het buitenland dan wel vanwege werk aan dit boek.

CURRICULUM VITAE

Gerrit van Oortmarssen werd in 1951 in Lochem geboren. In deze zelfde plaats haalde hij het HBS-B diploma, waarna hij van de Achterhoek naar de Technische Hogeschool in Twente verhuisde om Toegepaste Wiskunde te studeren. Hij haalde het ingenieursdiploma in 1977, na een doctoraalstage bij het instituut Maatschappelijke Gezondheidszorg en het Thoraxcentrum van de Medische Faculteit in Rotterdam. Sinds juli 1977 werkt hij op het instituut Maatschappelijke Gezondheidszorg, eerst op 3-maands contracten maar al spoedig op een meer bestendige basis, aan projecten op het thema "besluitvorming over bevolkingsonderzoek", onder meer van 1986-1990 op het samenwerkingsproject "Kosten en effecten van bevolkingsonderzoek op borstkanker" tezamen met Preventicon/Epidemiologie RU Utrecht en Epidemiologie/Radiologie KU Nijmegen. Hiernaast is hij betrokken bij andere projecten waarbij ziektemodellen ontwikkeld worden en bij de evaluatie van vroege opsporing van ziekten. In het bijzonder houdt hij zich vanaf 1985 bezig met ontwikkeling en toepassing van modellen voor evaluatie van de bestrijding van tropische infectieziekten. Zowel in 1993 als in 1994 vertoefde hij 3 maanden als *Guest Research Scholar* op het *International Institute for Applied Systems Analysis* in Laxenburg, Oostenrijk, voor onderzoek naar optimalisatiemethoden bij microsimulatiemodellen die de opsporing en bestrijding van ziekte beschrijven.

Hij is getrouwd met Fransien Kastanja en heeft twee dochters, Dagmar en Hainuwele.