

From the Department of Radiology, Stockholm Söder Hospital
Karolinska Institutet, Stockholm, Sweden

INCIDENCE AND INTERVAL BREAST CANCERS IN RETROSPECTIVE ASSESSMENT

Kerstin Moberg



Stockholm 2003

To my family

All previously published papers were reproduced with permission from the publisher.

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden
© Kerstin Moberg, 2003
ISBN 91-7349-573-5

ABSTRACT

The aims of this thesis were to investigate how different review methods used for radiological classification of breast cancers would affect the proportion judged to be false negative, to investigate how computer aided detection (CAD) would affect screening sensitivity, to study the balance between false positive and false negative results and to investigate patient outcome in relation to review result.

In a mass screening programme in Stockholm, commencing in 1989 and performed at five independent screening units, interval cancers (presenting in the time interval between two consecutive screening examinations) from two of the units, 103 women, and incidence cancers (diagnosed at subsequent screening examination) from one of the units, 117 women, diagnosed up until July 1993 were identified. Within a retrospective study set-up the screening mammograms, preceding breast cancer presentation, were reviewed mixed with images of healthy women at a ratio of 1:8 and non-mixed, including only cancer patients. Reviewers, both from the two units responsible for the screening mammograms and from two other units, participated in the study.

When using the mixed method, resembling the screen reading situation, significantly fewer cases were correctly selected (including very subtle findings) by the reviewers compared to when using the non-mixed method. Furthermore the number of reviewers (one, a majority or all) deemed necessary to classify a case to be detectable/false negative also influenced the rate that all in all varied between 7% and 34% for interval cancers and between 9% and 53% for incidence cancers. If cases correctly selected by a majority using either method were considering to be false negatives, the rate was found to be 22% for interval and 24% for incidence cases. Whether the reviewer worked at the unit responsible for the preceding examinations or not made no difference to the results. Trying to decrease false negative cases by increasing the number of women selected for recalls was in this retrospective study found likely to worry too many healthy women without providing a corresponding gain in cancer detection.

Preceding screening images of 58 women with interval breast cancer were reviewed mixed with other screening mammograms without and with the aid of Computer Aided Detection (CAD) and non-mixed by three radiologists. Despite sufficient sensitivity for CAD alone, equal to that of the reviewers, radiologists were not inclined to revise their original interpretations according to CAD and no increase in radiologist sensitivity, or decrease in specificity, was noted when using CAD. Specificity for CAD alone was low, 38%. Large scale prospective studies would be required in order to demonstrate the usefulness of radiologists using CAD in every day mass mammography screening.

No significant differences regarding tumour prognostic factors or survival could be demonstrated between different review categories, i.e. whether or not an abnormality was detected on preceding images. Reviewing of previous mammograms of women later diagnosed with breast cancer is an important tool in learning and teaching for both radiologists and radiographers. However there are better methods available for quality control of screening than the number of tumours retrospectively judged to be false negative.

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals:

- I. Moberg K, Grundström H, Törnberg S, Lundquist H, Svane G, Havervall L, Muren C
Two models for radiological reviewing of interval cancers.
J Med Screen 1999; 6: 35-39
- II. Moberg K, Grundström H, Lundquist H, Svane G, Havervall E, Muren C.
Radiological review of incidence breast cancers.
J Med Screen 2000; 7: 177-183
- III. Moberg K, Bjurstam N, Wilczek B, Rostgård L, Egge E, Muren C.
Computed assisted detection of interval breast cancers.
Eur J Radiol 2001; 39: 104-110
- IV. Moberg K, Blåsjö M, Sylvan M, Bjurling K, Sjöberg-Margolin S, Grundström H, Muren C. Pathological and clinical follow-up of radiological reviewed incidence breast cancers. In manuscript.
- V. Moberg K, Grundström H, Blåsjö M, Sylvan M, Muren C.
Pathological and clinical follow-up of radiological reviewed interval breast cancers.
In manuscript.

LIST OF ABBREVIATIONS

UICC	International Union against Cancer
WHO	World Health Organisation
NPI	Nottingham Prognostic Index
TNM	The UICC tumour classification based on tumour size (T), nodal status (N), and presence/absence of distant metastasis(M)
HRT	Hormone Replacement Therapy
CAD	Computer Aided Detection
CIS	Cancer In Situ

CONTENTS

1	Introduction.....	1
2	Background.....	2
2.1	Screening procedures and definitions.....	2
2.2	Factors limiting cancer detection.....	6
2.3	Reviewing the preceding mammogram.....	9
2.4	Computer Aided Detection (CAD).....	12
2.5	Review category - prognosis and characteristics	12
	Aims	14
3	Subjects and methods	15
3.1	Women included	15
3.2	Reviewing procedures.....	16
3.3	The computer aided detection (CAD) system.....	17
3.4	Primary treatment	18
3.5	Histopathology	18
3.6	Clinical follow-up	19
3.7	Radiological characteristics	19
3.8	Statistical methods.....	19
4	Results.....	21
4.1	Paper I - Review of interval cancers.....	21
4.2	Paper II - Review of incidence cancers	23
4.3	Paper III - Computer aided detection of interval cancer	26
4.4	Paper IV - Follow-up of incidence cancers	28
4.5	Paper V - Follow-up of interval cancers.....	35
4.6	Comparison between incidence and interval cancers.....	40
5	Discussion.....	42
5.1	Interval cancers (<i>I, V</i>).....	42
5.2	Computer Aided Detection (CAD) system (<i>III</i>).....	44
5.3	Incidence cancers (<i>II, IV</i>).....	45
5.4	General considerations	47
6	Conclusions.....	49
7	Acknowledgements	50
8	References.....	51

1 INTRODUCTION

Breast cancer is the most common malignancy and the most common cause of death from cancer in women in the western world. Breast screening with mammography at regular intervals offers a possibility to reduce mortality from the disease; through earlier detection the tumour will be stopped from progressing, allowing less extensive and more efficient treatment. In randomized trials a significant 30% reduction in mortality from breast cancer has been achieved [1].

Although mammography today is our best known tool for early detection of breast cancer no method can completely distinguish healthy women from diseased. Any diagnostic or screening test will unavoidably result in false positive and false negative cases. Generally the number of false positive healthy women is easily defined and accessible within a screening programme. On the contrary the number of false negative cases has to be evaluated by looking at the total number of women diagnosed with breast cancer after a screening test interpreted as negative. This evaluation requires a complete follow-up of all women who participated in the screening programme. It also includes a revision of the screening mammogram that preceded the breast cancer diagnosis. This retrospective review of the mammograms can be performed in different ways and as yet no consensus regarding which review method to be used has been established.

Regarding the extent to which the rate of false negatives depends on how the review was performed, the balance between false positive and false negative results, possible ways to limit the risk of false negative interpretations and patient outcome in false negative cases led to the research effort presented in this thesis.

2 BACKGROUND

2.1 SCREENING PROCEDURES AND DEFINITIONS

Breast cancer and prognostic factors

In the western world breast cancer is the most common malignant disease and the most common cause of death from cancer in women, with about 1 out of 10 women developing the disease during their life-time [2][3]. Differences in metastatic potential and survival, due to biological differences in tumour characteristics and growth rate, make breast cancer a heterogeneous disease.

The most important prognostic factors are tumour size and lymph node status at the time of diagnosis [4]. Tumour stage, as classified by the International Union against Cancer (UICC), in the TNM tumour stage, takes tumour size, regional lymph node status and presence or absence of distant metastases into account [5]. Histological classification as defined by the World Health Organisation (WHO) and histological tumour grade according to Bloom and Richardson modified by Elston and Ellis also provides prognostic information [6]. The Nottingham Prognostic Index (NPI) combines tumour size, nodal status and tumour grade [7].

The primary treatment of breast cancer is either breast conserving surgery followed by radiotherapy or mastectomy. Axillary clearance has been the standard procedure in patients with invasive tumours. Adjuvant systemic treatment is associated with a decreased risk of distant metastases and an increased likelihood of survival [8][9].

Service screening

As the technical development made radiological diagnosis of the disease in a pre-clinical phase possible, efforts were focused on mammography at repeated and regular intervals - mammography screening - as a tool to reduce mortality in breast cancer [10][11]. The aim of screening is to detect breast cancer at an early stage, during the *sojourn time* (Fig 1) preventing progress of the disease. This will allow less extensive (breast conserving surgery instead of mastectomy) and more efficient treatment, with the aim of mortality reduction.

Following the results of several randomised trials [12-24], one of them at Stockholm Söder Hospital [25][26], an overview of the Swedish trials in 1993 showed a significant reduction in breast cancer mortality of 24% in women aged 40-74 years and of 30% in women 50-69 years old [1]. Mortality reduction among women invited for screening compared to that of women not invited, instead of survival of the women diagnosed with breast cancer, was the endpoint of the trials to prevent *lead time bias* (this would occur if screening put the time of diagnosis forward without improving the outcome of the disease), Fig 1, *length bias* (slowly growing tumours with long pre-clinical phase are more likely to be detected by screening than fast-growing tumours) and *selection bias* (screening attenders probably being more observant and aware of breast cancers than non-attenders) from obscuring the results. The Swedish National Board of Health and Welfare 1986 recommended nation-wide mammographic screening [27] which was gradually introduced in the public health service of the Swedish counties and fully

effective by 1997. Although opinions still differ regarding the value of mammography screening the recommendations from 1986 have repeatedly been declared to remain valid. The so-called service screening programme invites the whole female population for screening regularly, in some counties women 40-74 years old; in others women 50-69 years old, as a rule every second year, often with 18 months interval for women ≤ 54 years old.

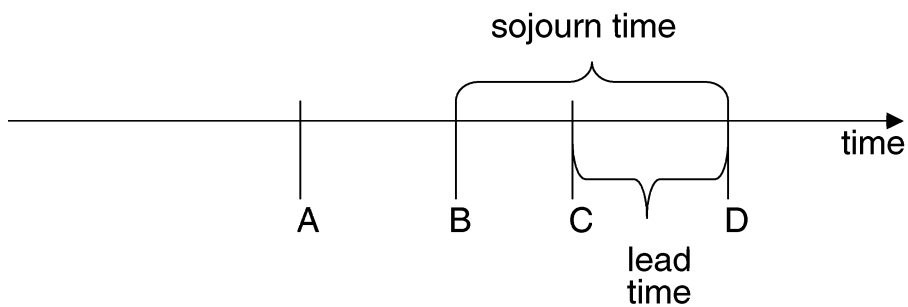


Fig 1 Diagrammatic presentation of sojourn time: time during which a lesion is potentially detectable in the preclinical phase(B-D) and lead time: time interval by which diagnosis has been brought forward by screening (C-D). A=biological onset of tumour. B=Tumour detectable by screening. C=Screening takes place. D=Clinical symptoms of cancer.

Both the absence of a non-invited control group and the many years of observation and follow-up required to assess the impact of the programme in terms of mortality reduction, have made early quality indicators important. The stipulated targets to be achieved in order to attain a reduction of the mortality rate [22], reported as having been obtained also in service screening programmes [28][29], are as follows:

1. The prevalent cancer detection rate should be three times the expected cancer incidence without screening;
2. > 50% of the screen-detected cancers should be < 15 mm in diameter;
3. > 70% of the screen-detected cancers should be node negative;
4. > 30% of the screen-detected grade III cancers should be < 15 mm in diameter;
5. The proportion of screen-detected grade III tumours of a given size should be equal to that expected without screening.

In a population regularly invited for screening breast malignancies can be diagnosed in women as: *Screen-detected* cancers as a result of the screening programme, *non-attender* cancers in women invited but not attending and *interval cancers* diagnosed between two consecutive screening rounds following a negative screen. The screen-detected cancers can be subdivided into *prevalent cancers* diagnosed at the first (prevalent) screening occasion and *incidence cancers* diagnosed at subsequent (incident) screening rounds, Fig 2.

A high *attendance rate*, at least 60% of those invited has been suggested as a minimum, is important for the impact of the programme as roughly between 30% and more than 50% of deaths among invited women have been reported to occur among non-attenders owing to the disease's high case fatality in this group of women [17][22][30][31].

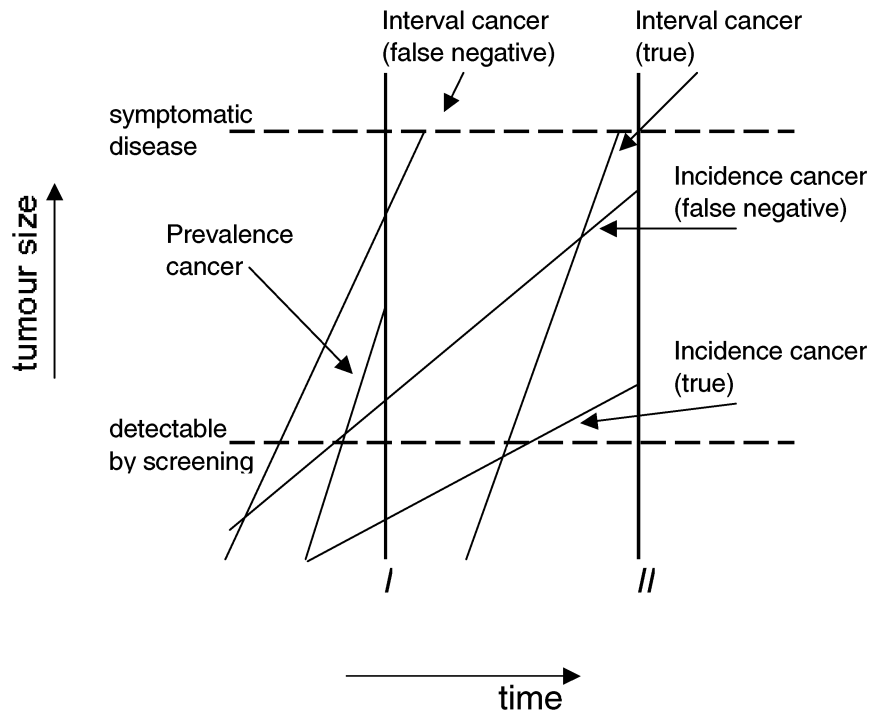


Fig 2 Diagrammatic presentation of prevalent(I) and incident (II) screening rounds. Cancers detected by screening and in the interval between screening rounds.

Rate of interval cancers has sometimes been considered to be a quality measurement of radiologists' performance [32] although other factors strongly influence these rates. The way interval cancers are defined and the methods used for their identification, the re-screening interval, the access of mammographic examination between screens, sojourn time and growth rate of the cancers, the way control cases are managed and the age distribution among those invited are some of these factors. As these factors vary between different screening programmes direct comparison of the interval cancer rates becomes inappropriate.

Women who attend screening and are found to have an abnormality raising suspicion of cancer are *recalled* for further assessment. The assessment includes a comprehensive diagnostic mammographic examination often complemented with ultrasound. If the suspicion of cancer remains the women will be *referred* for further diagnostic procedures and clinical assessment, sometimes including surgical biopsy. Reported recall rates has varied between 1.5% and 14.1% [33-35] while maximum rates of 7% for the prevalent and 5% for incidence rounds have been suggested [32].

As no diagnostic or screening method can completely distinguish between healthy and diseased women the screening procedure will unavoidable result in some healthy women being recalled and referred *false positive*, i.e. as diseased by the screening procedure while *specificity* is the proportion of the healthy women with a true-negative

result. On the other hand some diseased women will get a *false negative* result while the *sensitivity* of the screening procedure is the proportion of diseased women with a true-positive result. The sensitivity and specificity of a screening test are in many respects mutually exclusive, good quality implying their combination in a careful compromise.

Sensitivity is influenced by many factors; re-screening interval and age range of invited women, staff qualifications and quality of equipment, examination technique, recall and referral criteria and quality of the assessment procedure are some of these factors.

As the vast majority of the screened women are healthy even a minor reduction in *specificity* is important. High specificity, as well as sensitivity, is needed to provide an efficient and cheap tool for cancer detection and for women to have confidence in the programme.

False positive results in healthy women are known to create temporarily increased anxiety, sometimes creating a need for professional help, but rarely with a negative impact on the quality of the woman's future life [36-39]. To limit these adverse affects as much as possible all unnecessary medical interference, especially surgical biopsies with benign results, should be avoided.

False negative results in women with breast cancer by definition cause a delay in diagnosis which could potentially worsen the prognosis and also possibly lead to the semi-conscious suppression of an emerging palpable lump.

Interpreting screening mammograms

Interpretation of screening mammograms is a demanding task for the radiologist. The radiological appearance of normal breast tissue varies widely. While some women have breasts consisting mainly of low-density fatty tissue which is easy to interpret, women with dense, irregular, asymmetric, aberrant and hormonally activated breast parenchyma are not unusual. Breast cancer also has a varying radiological appearance. Tumours with a typically malignant appearance, such as spiculated lesions with dense centres and/or calcifications, will generally, even if small, be identified by a radiologist specialized in screening mammography. On the other hand tumours with only subtle signs of abnormality can be indistinguishable from ordinary parenchyma. Tumours simulating benign lesions like cysts are easily misinterpreted while some breast cancers are mammographically occult and can not be visualized at all.

In the Stockholm region, with approximately 1.8 million inhabitants, all women aged between 50 and 69 are invited every second year to one of the five independent screening units, somewhat more than 50.000 to Stockholm Söder Hospital. The screening radiologists thus make decisions upon hundreds of images every day within a limited period of time, with the one knowledge being that randomly mixed among the great majority of healthy women a few cases of malignancies might exist, as the number of tumours presented in a particular day or week is unknown. This is very different from a situation where patients are referred to the radiology department by clinicians because of symptoms and the radiologists know in whom and at which site to look for a lesion.

As screening involves an overwhelming majority of healthy women even a slight drop in specificity will cause many more healthy women to be recalled. For example in the Stockholm County 1 % lower specificity would imply the recall of approximately 1500 more women. On the other hand sensitivity has to be maintained at a level high enough to attain a reduction in mortality for women with breast cancer. Quality estimations in screening programmes must include both these groups of women. The demanding task for the radiologist is to detect as many tumours as possible while alarming as few healthy women as possible and especially to avoid unnecessary surgical interventions.

2.2 FACTORS LIMITING CANCER DETECTION

The radiologist's ability to detect a malignancy by mammography is a combination of the radiological characteristics of the lesion and the radiological characteristics of the ambient benign parenchyma, allowing the lesion to be distinguished from the surrounding tissue. As the quality of screen-film system, and in the future digital mammography, improves the threshold for cancer detectability will be lowered.

Technical limitations

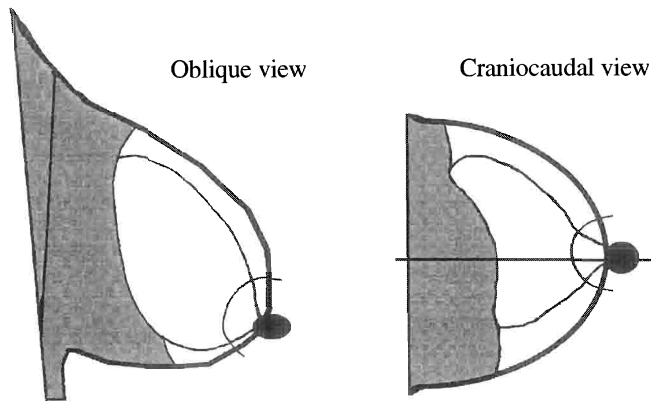
Of the factors limiting the detection or adequate characterisation of cancer on the mammogram the technical limitations in resolution and contrast associated with the screen-film system are among the most obvious. This is of particular importance in patients with dense parenchymal tissue. Parenchymal density is usually expressed according to Wolfe's classification in the N1, P1, P2 and DY categories, the categories expressing gradually increasing density. Dense parenchymal pattern constitutes a risk factor for breast cancer [40][41] and may obscure the detection of small, diffuse or low-density lesions.

Radiographic technique

An appropriate radiographic technique will result in well exposed images, with visualisation of a maximum amount of breast tissue at a minimum of radiation exposure. Proper compression of the breast during exposure will reduce scattered radiation and motion blurring, especially important in women with dense parenchyma. Proper positioning, with visualisation of the pectoral muscle and the inframammary fold, will allow the retroglandular tissue to be included in the images.

Lesion site

In the interpretation of screening mammograms Tabár and Dean [42][43] emphasise a viewing technique that will include those parts of the breast that require special attention, because of the high number of pathologic lesions arising in these locations, Fig 3. These parts of the breast are the retroglandular area, the retroareolar region and the medial half of the breast. As the retroglandular portion of the breast mainly consists of fatty tissue, dense parenchyma should not prevent lesion detection. However a tendency to overexpose the films slightly, in order to visualise dense tissue in the centre of the breast, may contribute to the difficulty in the perception of retroglandular lesions. Besides, dense tissue in the centre of breast may attract attention and perhaps cause a low-dense retroglandular lesion to go undetected, underlining the importance of a systematic approach to film reading. Digital mammography should offer ways of overcoming the problem of under/over-exposed areas in the image.



*Fig 3 Diagrammatic presentation of areas requiring special attention:
 Shaded region: The retroglandular area
 With semicircular and horizontal lines indicated areas: The retroareolar region
 and the medial half of breast*

Breast cancers show an uneven pattern of distribution within the breast with more tumours being located in the upper-outer than in the other quadrants [44]. A pattern of clustering of cancers has been demonstrated in the axillary tail and inframammary fold on the oblique projection [45]. On the craniocaudal projection clusters were noted in the tail, central, dorsal and medial portions of the breast. The site of the lesion can make it difficult to visualise in more than one of the two standard views (Fig 3), producing a false belief that the lesion is not real [46]. Unlike the technical limitations of the screen-film system, poor radiographic technique can be improved by measures undertaken by the staff; their attention being attracted by the system providing continuous quality assurance [47-49].

Lesion characteristics

The radiological characteristics of the lesion can make it more or less difficult to distinguish from the ambient parenchyma. Ultrasound is a useful complement in doubtful cases but requires that the woman be recalled for further assessment. The sonographic finding must also be convincingly correlated to the mammographical abnormality in order to dismiss cancer suspicion.

Small lesion size, with the exception of branching calcifications, lack of signs of malignancy relating to tissue fibrosis, such as spiculations, parenchymal deformities and nipple/skin retractions, complicate lesion detection [50-53]. Invasive lobular carcinoma can lack desmoplastic reactions with an asymmetric density being the only sign of abnormality [54]. Other non-characteristic subtle signs are architectural distortions and developing unspecific parenchymlike densities.

Improving radiologists' sensitivity to these subtle signs of abnormality has been suggested as a way to increase cancer detection rates and decrease interval cancer rates

[50]. Although breast cancers are often preceded by such subtle or minimal signs [55-57], no additional cancer risk, compared to that expected in asymptomatic women, has been found for women with minimal signs [58][59]. From this follows that increased radiologist sensitivity to minimal signs must be balanced against an expected decrease in specificity to avoid worrying too many healthy women with unnecessary recalls. Nevertheless it is also evident that these subtle signs often precede more evident signs of malignancy and that this change in radiological appearance reflects and is highly consistent with the natural history of the disease [60][61]. In other words the disease is more easily detected when prominent than when subtle.

On the other hand, malignancies of some histological subtypes are associated with a detectable radiological appearance already when small: for example tubular carcinomas usually have the appearance of easily detectable spiculated lesions [62-63]. This phenomenon has given rise to reflections concerning possible detection bias, implying that mammography more often will detect lesions with a good prognosis than those with a worse outcome. However, common radiological signs of malignancy, like spiculations and densities, are associated with different histological categories, including tumours with high malignant potential [60-61][64]. The presence of casting calcifications are regarded as an important prognostic factor for small invasive cancers [65][66].

Related to the difficulty of detecting early subtle signs of malignancy is the difficulty in communicating these findings. Considerable variability in the characterisation of mammographic abnormalities has been reported [67] even for such an undisputable sign of malignancy as "spiculation"[68]. Digital mammography offers a possibility for introducing a perceptual aid (see 2:4), but may also facilitate uniform terminology for lesion characterisation and improve communication of these findings.

Interpretation errors

A systematic approach to each mammogram, optimal viewing conditions and avoidance of all unnecessary interruptions and distractions are important for the minimization of interpretational errors [69]. Radiologist's nonbelief, fatigue, inattention and inexperience are factors influencing interpretational skill. Misinterpretation may sometimes result from the discovery of an obvious finding leading to a more subtle lesion being overlooked. Slow or no apparent tumour growth could be misinterpreted as a benign finding or a normal breast pattern due to the absence of interval change [70-71]. Some lesions that turn out to be malignant have a mammographic appearance suggesting that they are benign, for example mucinous carcinomas can have an appearance resembling that of benign cysts. Clinical complaints or findings noted by the woman herself or by the radiographer, as well as the location of any surgical scars, should be noted. Postsurgical changes can obscure the lesion or mislead the radiologist and should either be stable or decrease with time [71].

Reported ways to increase sensitivity with essentially preserved specificity are: double reading with consensus recall decisions[72-74], two-view instead of one-view examinations [74-75] and re-screening intervals not exceeding 2 years or preferable shorter for women between 40 and 55 years old [22][76].

The term false negative can be used to describe all examinations where a breast cancer is present in spite of a negative mammogram. In review studies, as in this thesis, the term is used to signify cases where an abnormality suggestive of malignancy could be detected on retrospective review of the preceding screen images.

2.3 REVIEWING THE PRECEDING MAMMOGRAM

Review categories

Identifying a case as overlooked or misinterpreted is admittedly a subjective procedure. It is based on a review where images performed at breast cancer presentation are compared to the previous screen images and classified, in accordance with suggested guidelines, into the following categories [32]. If the tumour is undetectable at the preceding screen the case is classified as *true*. When the tumour is undetectable both at previous screen and at presentation a case is classified as mammographically *occult*. In cases overlooked or misinterpreted due to *observer error*, also called missed or false negative cases, an abnormality suggestive of malignancy is identified at the site of the presenting tumour on previous screen images. When revision shows only a subtle abnormality, the appearance of which suggests low probability of malignancy, the case is classified as *minimal sign* present. Cases missing images from cancer presentation are *unclassifiable*. A breast cancer that has been overlooked or misinterpreted could later be diagnosed as an incidence cancer, detected at the next screening round, or it may be detected during the time interval before next screen as an interval cancer (Fig 2). In women who choose not to attend further screens the tumour could be detected later among non-attenders.

Factors influencing review results

Whenever a breast cancer is diagnosed and the preceding screen images are available a review generally takes place. As presenting images have usually already been interpreted, the radiologist now knows "what kind of abnormality, at what site, in which patient". As the diagnostic examination may take place at the same unit responsible for the preceding screen, both the original interpretation and the review may be carried out by the same radiologist. Thus basic conditions certainly differ between the original screen reading and the retrospective review.

In applying a review situation which imitates that of the original screen reading, the ideal would be to randomly include the preceding images of cancer patients with current ordinary screening mammograms, preferably at another screening unit using external reviewers. However the rapid development of technical equipment would reveal an examination which was only a few years old to be of an earlier date than the other mammograms and patient identification, date and hospital markings would indicate that the mammograms originate from another screening unit.

A number of factors, apart from the interpretational skill of the original screen reader, could potentially influence the proportion of cancers deemed to be detectable on retrospective review of preceding screen mammograms. Differences in the reviewers' reactions to the review study situation, in itself differing from ordinary screening conditions, could influence their retrospective detection rate. Whether the reviewer is blinded to patient outcome or not and whether he/she has access to images performed at cancer presentation or not has been shown to have an effect [77]. Studies indicate that if

images of healthy women are included in the review set, randomly mixed with images of cancer patients, fewer cancer cases will be detectable on review [78][79]. The proportions of cancer patients and healthy women in the mixed review set would hence also affect detection rate.

Differences in the number of reviewers deemed necessary in order to classify a case as detectable (all of them, a majority or just one of them) will influence review results [80][81]. It is unclear whether or not the inclusion of external reviewers (not working at the screening unit responsible for the original screen interpretation) would make a difference. Comparing the rate of cancers classified as detectable on retrospective review from different studies is thus difficult as the rate is influenced by many factors apart from the skill of the original interpreter. Yet no consensus has been established as to what method should be used. Besides including a high number of unclassifiable cases would make the proportion presented as false negative to seem smaller.

Review studies

Most review studies have so far concerned interval cancers, sometimes providing few details as to how the cancers were identified and to how the review procedure was actually performed. Generally the preceding screen images have been reviewed separately, non-mixed, with reviewers aware of patient outcome (the later diagnosed interval cancer) and having access to presenting mammograms and sometimes by radiologists involved in the original interpretation. The proportion of interval breast cancers regarded as observer error (not including the "minimal sign" cases) in those studies varied between 6% and 43% [82-97]. Reviewing 103 incidence cancers non-mixed, but without access to presenting mammograms, 31% were found to be false negative [98]. Studies including both interval and incidence cancers have found the proportion of detectable incidence cancers to be equal to or higher than that of interval cancers, for example in a Finnish study 43% versus 19%, [99][55].

A few studies have *compared different review methods*. When reviewing 73 women with palpable breast cancer non-mixed, without access to presenting mammograms, 41 % were found to be false negative. An additional 34 % of the remaining 43 cases were considered missed when reviewed with access to presenting mammograms [77]. In another study using similar methods for incidence cancers 25% were found to be detectable on blind review and an additional 19% on informed review [100].

The rate classified as missed cases varied between 4 % and 56 % when 50 interval cancers in a study were reviewed mixed with normal mammograms at a ratio of 1:3 and non-mixed [78]. The rate depended both on the review method used (mixed or non-mixed) and on the number of reviewers (all six or just one of them) deemed necessary to classify a case as false negative.

In some studies *only a mixed review method* has been used. Approximately 70% of invasive interval cancers were selected for further assessment by four out of five radiologists when 106 screening mammograms were reviewed using a mixed method by Day et al [101]. The mixed set comprised one third normal cases, one third screen-detected cancers and one third interval cancers. However no correlation was made between the location of the selected lesions and the presenting tumours.

In a study reviewing incidence cancers mixed with normal screens at a ratio 1:3, 19% (21/112) were found to be missed cases on the basis of majority consensus among three reviewers. In a later session the cases not classified to be missed were further divided into "minimal signs", 32 (28 %), or "new" tumours [79]. Using the same mixing proportions in another study 23% (30/133) of the reviewed incidence cancers were regarded as missed cases if correctly identified by at least two out of four external radiologists [102].

Some studies have focused on the *number of reviewers* deemed necessary in order to classify a case as false negative. An unblinded non-mixed review method applied by the East Anglian region in the UK classifies a case as false negative according to a majority consensus decision among radiologists representing the seven screening units in the region. Using this classification 14% of reviewed interval cancers were found false negative, while the minimum rate (all reviewers agreed to classify a case as false negative) was 6% and the maximum rate (only one reviewer needed to classify a case as false negative) was found to be 38% [80].

The reviewers in a study using a similar method, regarding a case as missed if correctly identified by at least two of the participating reviewers within a group of on average seven radiologists meeting regularly, classified 26% of interval cancers as false negative [103].

Using mixing proportions of 4:1 a study found the false negative interval cancer rate to vary between 32% and 22% dependant upon whether two or four of the eight participating reviewers agreed to identify a case as missed. Forty percent of the normal mammograms included were selected for recall by at least two reviewers in the same study [81]. Using a majority consensus procedure has the advantage of being independent of the number of participating reviewers and probably therefore more reproducible.

A *two-stage method* has been reported: interval cancers from seven screening units, 129 cases, were in a first stage reviewed mixed according to the proportions described by Day [101]. A number of 38 % were found to have been correctly selected by a majority of reviewers. In a later unblinded informed session 41% were found to be false negative and a further 16% to have minimal signs by majority consensus decision among the same reviewers [104].

A similar method with the difference being that the preceding interval cancer images in the first step were included in the daily workload at the screening units has been described [105]. In that first stage 15% were selected for recall, in the second informed stage 25% were found false negative and a further 24% as showing minimal signs by majority consensus.

Reviewing 54 incidence cancers mixed with normal screening mammograms at a ratio of 1:3, followed by an informed majority consensus decision among three radiologist, 26% were found to be false negative [106].

A two-stage method has the advantage of offering both a tool for estimating the proportion of cancers possible to detect on earlier images (blinded review) and a tool for continuing the education of screening radiologists and radiographers to improve screening quality (consensus discussion).

2.4 COMPUTER AIDED DETECTION (CAD)

Double reading can increase sensitivity by increasing cancer detection rate and by reducing the number of false negative cases [107][72][73]. Double reading implies that the images are independently interpreted by two radiologists, this is then usually followed by a consensus decision as to whether a woman is to be recalled or not. However, double reading could be difficult to achieve as the number of screening radiologists is limited.

Systems for Computer Aided Detection (CAD) are rapidly developing and might be useful as a perceptual aid for radiologists reading screening mammograms [108][109] and could be integrated in future digital mammography systems. These systems, as the Image Checker M 1000 System (R2 Technology Inc, Los Altos, Ca, USA), are primarily optimised for tumours with typical signs of malignancies, spiculations and calcifications. Earlier studies concerning tumours with these characteristics have shown that CAD can increase radiologists' sensitivity in the detection of spiculated lesions [110] and clustered microcalcifications [111].

Two radiologists were found to have a modest increase in sensitivity with unaffected specificity when using CAD in a study by Thurfjell et al. [112], although the specificity of the CAD system alone was very low. The study included 64 screen-detected cancers, 10 interval cancers and 46 normal healthy controls, 120 screening mammograms in total. The CAD system, despite showing satisfactory overall sensitivity, failed to detect any of the 10 interval cancers, but even an expert screener detected only 2 of them.

CAD prompted 77% (88/115) of lesions judged as actionable (warranting recall) in a study comprising 427 available prior mammograms of 1083 women later diagnosed with breast cancer [113][114]. Based on the same set of 1083 cancer patients, CAD correctly marked 86% (322/ 375) of those cancers which were judged by 3 radiologists to be spiculated lesions on diagnostic images [68]. In a third and separate study, based on the subset of lobular cancers among the same 1083 malignancies, CAD managed to correctly mark 91% (86/ 94) on diagnostic images and 77% (24/ 31) on available prior screening mammograms [115].

In a prospective study based on 12 860 screening mammograms performed during one year and interpreted with and without the CAD-system a 19.5% increase in cancer detection rate (41 cancers detected without CAD and 49 with CAD) was reported [116]. Recall rate increased from 6.5% to 7.7% (from 830 to 986 = 156), although more than 97 % of all CAD computer marks were dismissed by the radiologist. Of the 156 additionally recalled women 21 had a surgical biopsy, 8 with a malignancy diagnosed. Of the 8 additional tumours found with CAD histology showed 5 cancers in situ (CIS) at stage 0 and 3 invasive cancers at stage I.

2.5 REVIEW CATEGORY - PROGNOSIS AND CHARACTERISTICS

Patient outcome and prognostic factors

No significant differences in breast cancer mortality were found between different review categories of interval cancer patients in studies of 544, 191 and 75 interval cancers respectively [117][88][93] In the largest of these studies false negative cases

had significantly higher proportion of grade I tumours than true interval cancers [117]. Prognostic factors did not differ significantly between review categories in two other studies of 385 and 90 interval cancers respectively [118][95]. Interval cancers constitute a heterogeneous group of tumours including a subset of fast growing cancers among true interval cancers [93][119-121].

Few follow-up studies of differences between review categories of incidence cancers have been reported so far. In studies including 112, 133 and 103 incidence cancers respectively no significant differences in prognostic factors between tumours of different review categories were found [79][102][98]. In two studies, including 76 and 54 incidence cancers respectively, false negative tumours were found to be of lower tumour grade, but without differences in tumour stage [99], and with more favourably NPI [106] compared to true incidence cancers.

Parenchymal pattern, lesion site and appearance, recall rate

Studies comparing parenchymal density of different groups of tumours have shown dense parenchymal pattern to be associated with increased risk of interval versus screen-detected cancer [122-125], of tumours of high grade versus tumours of low grade [123] and of occult cancers [117][118][127]. Dense parenchyma is a risk factor for women with a family history of breast cancer [126] while results differ whether parenchymal density also is a risk factor for false negative tumours [70][122].

False negative incidence cancers tended to more often be situated in any of the areas requiring special attention (fig 3) than true cases in a study [100]. Deep retroglandular lesions accounted for one third of missed cancers in two studies [70][45], however, almost the same proportion of the screen-detected tumours were located in the retroglandular region in one of them [45]. Only 37% (7/19) of subareolar lesions, where dense tissue often is present, were mammographically detected in another study [127]. Missed cancers were also found more likely to be visible in only one view [70][46].

Comedo-type calcifications have low probability of being missed [50][51] while lesions lacking signs of malignancy related to fibrosis were judged to contribute to 22% of false negative cases [70]. Invasive lobular carcinomas are common among occult tumours [117-118]. Subtle signs of malignancy as architectural distortion, asymmetries and developing unspecific parenchymalike densities are reported more likely to be missed [50-53]. However, false negative tumours have not been found to have any specific radiological appearance at presentation differing from true negative cancers [117-118].

Recall rate at preceding screen have reported higher, in one study 7%, among interval cancer patients compared to the rate in the general screened population, 5% in the same study [118]. Recall rate was found especially high in the false negative group [118][128-129].

AIMS

- To investigate whether the numbers of interval/incidence cancers with abnormalities possible to detect in retrospect would differ if the review procedure resembled the screening situation compared to a review procedure including only the cancer cases.
- To investigate how differences in cancer detection rate in a study situation, especially of tumours at an advanced stage at presentation, would influence the number of healthy women selected for recalls
- To investigate whether study detection rate would differ if the reviewer worked at the screening unit responsible for the reviewed mammograms or at another unit.
- To investigate how the proportion of tumours regarded as false negative would vary depending on the number of reviewers (just one, a majority, or all) considered necessary to classify a case as missed
- To investigate both the interval cancer detection rate of a CAD system compared to that of radiologists in a randomly mixed review situation and how radiologists' sensitivity and specificity would be influenced by CAD.
- To investigate whether tumour characteristics, prognostic factors and patient outcome differ between interval/incidence cancers of different review categories (i.e. whether retrospectively detected or not).

3 SUBJECTS AND METHODS

3.1 WOMEN INCLUDED

In the Stockholm county region, an area with now approximately 1.8 million inhabitants, a service screening programme, previously described in detail [33], started in 1989. Women between 50 and 69 years old are invited every second year to one of the five independent screening units (A, B, C, D and E). At one of the units, Stockholm Söder Hospital, in this thesis referred to as unit D, the mass screening programme was preceded by a randomised trial [25][26] immediately followed by an invitation of women from the control group for screening. In women examined for their first (prevalent) time two views are regularly taken whereupon the radiologists decide whether one or two views are to be taken at the next (incidence) screen.

From July 1 1989 until June 30 1991, a total of 107 846 women, 70.6 % of those invited, attended screening in the region. Breast cancer was diagnosed in 756 attending (0.7 %) and in 252 non-attending women (0.56%).

Interval and incidence cancers were identified by the Regional Oncologic Centre, Interval cancers by matching the screening and cancer registers, both based on personal ID-numbers. A number of 207 women had a breast cancer diagnosed within two years of a negative first round screen, giving an interval cancer rate of 1.9 / 1000 women screened.

In 60 of those 207 cases the preceding screen was performed at screening unit D (Stockholm Söder Hospital) and in 44 cases at unit C (Danderyd Hospital). One interval patient from unit D was excluded due to reappraisal of the cancer diagnosis with withdrawal of the cancer registration.

From the screening register 119 women were identified with screen-detected breast cancer at unit D before June 30 1993 and previous round attendance within the programme (incidence cancer). Screening images were available for 117 of these women. Figure 4 shows women included in paper I-V.

Papers I and V

The 59 interval cancer patients from unit D and the 44 from unit C were included. Mammograms of 390 healthy women, of 15 women with screen-detected breast cancer and of 11 women referred because of lesions proved benign, were included in the radiological review. In one interval cancer patient the screening images were missing, she was excluded from the radiological review but included in the survival analyses. Included in paper I were thus mammograms representing 518 women. Paper V included the 103 interval cancer patients of whom one had bilateral malignancies.

Paper III

The 59 interval cancer patients from unit D, of whom 58 with available screen images, and the mammograms of 211 healthy women, 5 women with screen-detected breast cancer and 5 women referred because of lesions proved benign were included in the review set thus representing 279 women in total.

Papers II and IV

The screening mammograms of 117 incidence cancer patients together with images of 769 healthy women, 29 women with screen-detected breast cancer and 21 women referred because of lesions found to be benign were included in paper II. The review set thus represented 936 women. Paper IV included the 117 incidence cancer patients of whom five had bilateral malignancies.

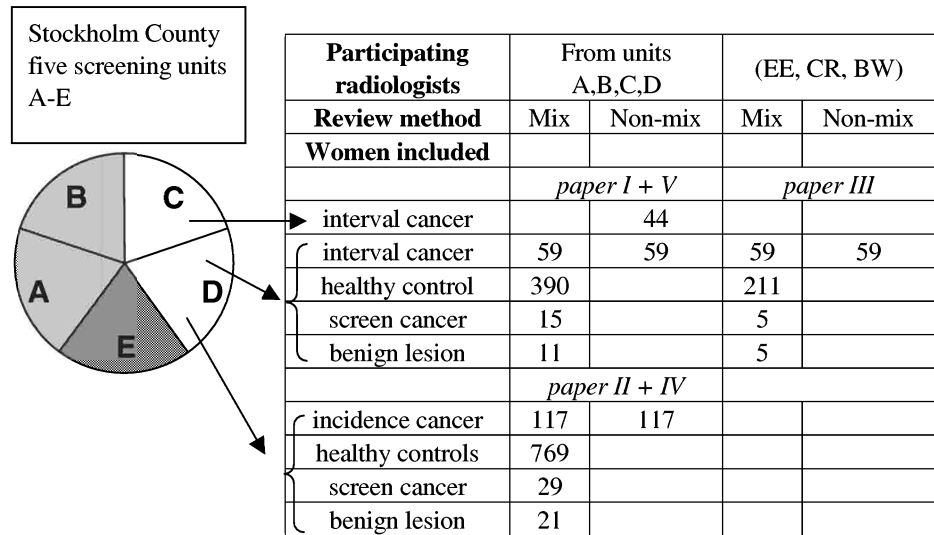


Fig 4 Women and reviewers included in paper I-V.

3.2 REVIEWING PROCEDURES

In paper I and II eight radiologists participated from the four screening units A, B, C and D while in paper III three radiologists (EE, CR, BW) took part.

Mixed review (papers I-III)

To constitute the mixed review set screening images preceding the interval (I and III) or incidence (II) cancer diagnosis at unit D were mixed with the screening mammograms from the other included women, ratio 1:8 (I and II) and ratio 1:4.8 (III), using a computer-generated random number sequence.

The reviewers recorded the findings by filling out a standardised follow-up form, indicating the location of any abnormality and whether it was suspect enough to motivate recall of the woman for further assessment or not.

Non-mixed review (papers I-III)

In a later session the interval cancer cases, including the 44 cases from unit C (I), and the incidence cases were reviewed separately, non-mixed, the reviewers now knew that all examinations represented women with a subsequent malignancy but had no access to the diagnostic mammograms performed at presentation.

If a case was selected to be recalled and the site of the indicated abnormality was identical to the later diagnosed incidence or interval cancer (or to the screen-detected

cancer or benign lesion) the case was classified as *correctly* selected. The correctly selected interval and incidence cases were in some analyses further divided into those selected *correctly by a majority* (>50%) of reviewing units in mixed or non-mixed readings and those selected *correctly by a minority*. Interval and incidence cancers not selected or selected for the *wrong area* (wrong area or the opposite breast) were summarized as *not correctly* selected. These radiological categories (Not correct, correct by majority and correct by minority) are referred to as the *review category* of the case.

In study III the images were read independently by the three radiologists, first without and then with aid of CAD. In study I the mixed reading was performed by unit A, B and C, one reviewer from each unit while the mixed reading in study II as well as the non-mixed review in both studies was performed by all four units, two radiologists from each unit. In study two the reviewers filled in separate forms to evaluate single- and double-reading and in study III separate forms for reading with and without aid of CAD.

3.3 THE COMPUTER AIDED DETECTION (CAD) SYSTEM

The CAD system (Image Checker M 1000, R2 Technology Inch, soft ware version 1.2 at the time of the study) is optimised to detect and mark "regions of interest" (ROI), acting as a perceptual aid to radiologists reading screening mammograms. After digitizing the image in a laser scanner the CAD analyzing system searches for and marks clusters of calcifications with a triangle and spiculated masses with asterisks, Fig 5. The system is optimised for masses of size 10 to 20 mm. The digitised images, with ROI marked, are presented on two low resolution 13 x 13 cm video-monitors built into a multiviewer.

Using CAD the radiologist first makes a conventional interpretation of the mammogram. With the digitized images available on the video monitors, with possible malignant areas indicated, the radiologist re-examines the original mammogram. The equipment thus allows the radiologist to re-examine the mammogram while viewing the image analyzing system result and decide whether to revise the original interpretation or not.

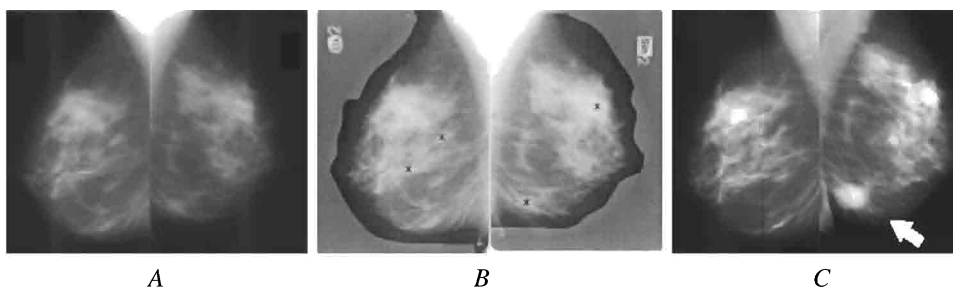


Fig 5 A Screening mammogram interpreted as normal. B Same examination as A with CAD markings in both breasts. C Mammogram at presentation with interval cancer in left breast indicated by white arrow.

3.4 PRIMARY TREATMENT

The primary management of patients with detected breast cancer was carried out in accordance with guidelines published by the Stockholm Breast Cancer Study Group [130]. Generally breast conserving surgery with axillary dissection was offered to patients with stage I and stage II tumours less than 30 mm in size followed by postoperative radiotherapy to the breast. In more advanced cases the women were treated by mastectomy with axillary dissection. Adjuvant systemic therapies were offered according to the guidelines.

Women with incidence cancer

Out of the 117 women with incident breast cancer two (one with bilateral malignancies) had neoadjuvant chemotherapy in order to decrease breast tumour mass before surgery. A number of 48 (41%) women, five with bilateral malignancies, were surgically treated with mastectomy and 69 (59 %) with breast conserving surgery. Only eight women (7 %) had noted any clinical symptoms in relation to the malignancy at presentation.

Women with interval breast cancer

Out of the 103 women with interval breast cancer (102 of them radiologically reviewed) two were not surgically treated, one due to stage IV disease already at detection while the other was in too poor general condition. Four women were treated with neoadjuvant chemotherapy, three of them in order to decrease breast tumour mass before surgery; in the fourth case, an elderly woman with a very slowly progressing tumour, tamoxifen was successfully given but surgery was nevertheless carried out 1.5 years later. Of the 101 surgically treated women (one with bilateral malignancies) 62 received mastectomy and 39 had breast conserving surgery. Of the 94 invasive and surgically removed interval cancers the procedure included axillary dissection in 91 cases (90 of them radiologically reviewed), while in 2 women, with small infiltrative components (1 and 3 mm respectively) no axillary dissection was carried out. In 16 (16%) of the 103 women the tumour was detected mammographically without the patient noticing any clinical symptoms of malignancy while 87 (84%) women were referred to mammography because of clinically suspicious symptoms of malignancy.

3.5 HISTOPATHOLOGY

Data on tumour size, measured as the widest diameter in the specimens, histological classification as defined by the WHO [131] modified by Linell [132] and pTNM classified as defined by the UICC [5] were taken from the original pathological reports. In cases of predominantly cancer in situ where infiltrating components were also present, the size of the invasive components were noted, if microscopic invasion was reported without any size given, the size was considered 1mm.

For invasive cancers histological tumour grade, using the Elston-Ellis modification of the Bloom-Richardsson method [6] was established retrospectively and independently by two pathologists (MB, MS) blinded to the original report as well as to outcomes of the radiological review and clinical follow-up. The case was re-reviewed by both pathologists when their opinions differed and then reported in consensus.

3.6 CLINICAL FOLLOW-UP

The Causes of Death registry is a nation-wide report registry covering Swedish residents and linked to population statistics. From that registry data was retrieved for every woman with interval or incidence cancer included in the study. It was thus established whether the woman was dead or alive at the end of the observation period 1 January 2001 and what the reported cause of her death was. When a cause-of-death certificate is missing a request is sent out to the medical establishment and the drop out rate is therefore reported to be very low (less than 0,5 % of all deaths). Women with incidence or interval cancer in this study were included in the breast cancer data base of the Stockholm Breast Cancer Study Group for the Stockholm-Gotland region. Follow-up data in the register is based on continuous reports from the clinicians responsible. Data on presence of distant metastasis, local recurrences and contralateral breast cancer up until 1 January 2001 was retrieved from this same data base and complemented with records from the mammography, pathology and oncology department at Stockholm Söder Hospital/ Huddinge University Hospital and Danderyd Hospital.

3.7 RADIOLOGICAL CHARACTERISTICS

The general density of the breast parenchyma on the preceding screen images was classified into four subgroups N1, P1, P2 and DY according to the Wolfe system [40-41] by an expert breast radiologist (KB) not previously involved in the project and blinded to the review results, pathological reports and clinical outcome.

The location of the incidence or interval cancer was thereafter marked on the (preceding screen) images whereupon the same breast radiologist classified the parenchymal density in the tumour area according to Wolfe as described above. Whether the marked location was situated in any of the areas requiring special attention (Fig 3) or not, as well as the radiological appearance of the correctly selected lesions was evaluated by the same radiologist.

Radiological appearance was classified in the following categories: calcifications only, spiculated mass (with or without calcifications), asymmetric density, parenchymal distortion, rounded mass (with or without calcifications), unspecific parenchymlike density (with or without calcifications) and unclassifiable (too subtle to be classified).

The radiological tumour appearances at presentation were scrutinised and classified in the same categories by a breast radiologist (CM) blinded to case review category, pathological report and patient outcome. The rate of preceding screen assessments and possible reasons for the false negative cases (correctly selected by majority) being overlooked/misinterpreted were then evaluated (CM, HG, KM).

3.8 STATISTICAL METHODS

To compare the rate of correctly selected cases between the mixed and the non-mixed review methods the Wilcoxon Matched Pairs Test and McNemar's test (I, II) was used.

To compare the rate of correctly selected women and specificity between reviewing units (I, II) Cochran's Q Test was utilized.

Independence test of contingency, Chi-square tests, asymptomatic or exact (Fisher's exact test), were performed to compare nominal or ordinal data for assessment of association between review category (correctly versus not correctly selected malignancies) and tumour characteristics (IV, V).

To compare the means of age, time interval and so on, between review category groups, one-way Anova and t-tests were used.

The non-parametric Mann-Whitney U-test and Kruskal Wallis Anova by ranks was utilized for assessment of differences in continuous data (such as tumour size) followed by multiple comparisons between review category groups based on mean ranks.

For survival analysis the life-table method according to Kaplan-Meier was used [133-134]. Women with breast cancers who died of other causes were censored, as were women with contralateral invasive breast cancer from the date of diagnosis. Two-tailed p-values < 0.05 were considered statistically significant.

The studies were approved by the ethical committee at Karolinska Institutet, Stockholm, Sweden.

4 RESULTS

4.1 PAPER I - REVIEW OF INTERVAL CANCERS

Mixed review

The results from the mixed review are summarized in Table I. Although the frequency of women selected to be recalled for further assessment due to mammographic findings varied between the reviewers, the number of correctly selected women with later diagnosed interval cancer was at almost the same level, unit A 9/58 (16%), unit B 7/58 (12%) and unit C 9/58 (16%).

Reviewing screening unit	Women with interval cancers		Healthy women	
	Correct	Wrong area	Selected	Specificity
	No. (%)	No. (%)	No. (%)	%
A	9 (16)	6 (10)	41 (11)	89
B	7 (12)	1 (2)	11 (3)	97
C	9 (16)	6 (10)	32 (8)	92

Table I Mixed review: Women correctly selected and those selected for the wrong area among women with interval cancer (n=58). Specificity and number of selected among healthy women (n=390).

Units A and C selected almost as many interval cases for the wrong area (wrong area or opposite breast) as they selected correctly (6 versus 9 women) while unit B selected only one case for the wrong area. Among the screening mammograms representing the 390 healthy women specificity was 89%, 97% and 92% for unit A, B and C respectively.

In other words a high recall selection rate mainly increased the number of selected women who either did not develop an interval cancer (healthy women) or who developed an interval cancer at another site than the one selected or in the opposite breast (women with interval cancers selected for the wrong area).

Non-mixed review

The results from the non-mixed review are summarized in Table II and the whole review procedure in figure 6. The number of correctly selected women with interval cancer from unit D increased significantly and was almost twice the number at the mixed review (Wilcoxon Matched Pairs Test $p = < 0.001$ and Mc Nemar's test $p = 0.0001$).

Unit A selected almost as many interval cases for the wrong area or opposite breast (18+12=30) as it selected correctly (16+15=31), i.e. in only 50% of women with reviewed interval cancer selected by unit A was the indicated abnormality identical with the lesion later diagnosed as the interval cancer. Unit B did not select any for the wrong area, which is in accordance with the high specificity for unit B (Table II). Unit C and D had a small number of cases selected for the wrong area.

Reviewing screening unit	Correct				Wrong area			
	from unit D		from unit C		from unit D		from unit C	
	No.	(%)	No.	(%)	No.	(%)	No.	(%)
A	16	(28)	15	(34)	18	(31)	12	(27)
B	15	(26)	10	(23)	0	(0)	0	(0)
C	16	(28)	11	(25)	3	(5)	2	(5)
D	11	(19)	7	(16)	1	(2)	3	(7)

Table II Non-mixed review: Correctly selected and women selected for the wrong area among women with interval cancer from unit D (n=58) and unit C (n=44).

Mixed and non-mixed review procedure

Comparing selection rate for the participating reviewing units, it remained unaltered and independent of whether the unit was responsible for the reviewed preceding screen examination (unit C and D), or not, Fig 6.

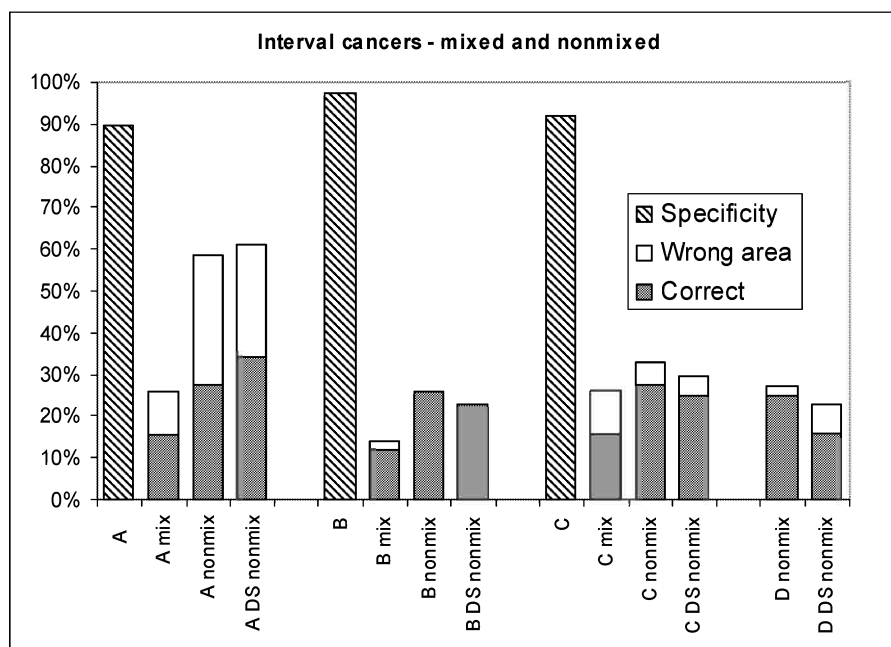


FIG 6 Selection rate, correctly and for the wrong area, of women with interval cancer. Cases from unit D (n=58) reviewed by the mixed and non-mixed method, from unit C (=DS, n=44) by the non-mixed method. Specificity for healthy women (n=390).

Agreement among the reviewing units as to which women with interval cancer who should, or should not, be selected by the two methods is shown in fig 7 a and b. Preceding mammograms from 4 of the 58 women with interval cancer from unit D were correctly selected by all screening units using both review methods. These 4 women (7%) constitute the minimum number who could be considered as missed cases

due to observer error. On the other hand 20 (34%) were correctly selected by at least one of the reviewing units using either of the two review methods. These 20 women (34%) constitute the maximum number that could be regarded as missed cases. The majority, 78% and 66% respectively, of the interval cancers were not correctly selected by any of the reviewing units using either of the two review methods (Fig 7 a and b).

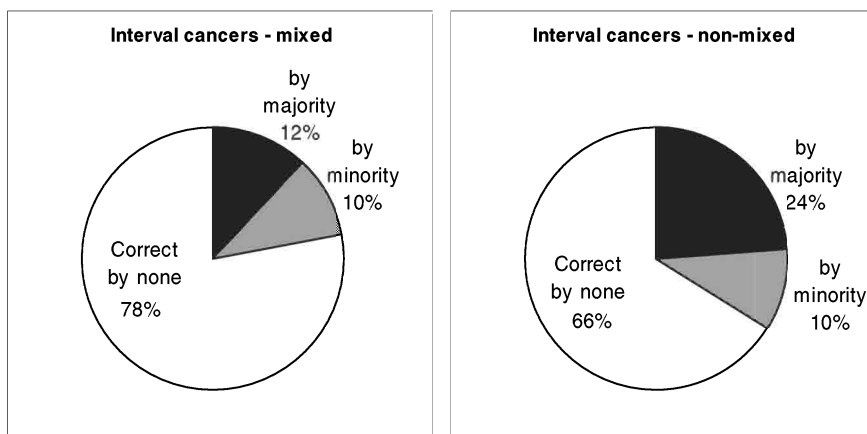


FIG 7 Interval cancers reviewed mixed and non-mixed (n=58) correctly selected by none, a majority or a minority of the participating screening units.

4.2 PAPER II - REVIEW OF INCIDENCE CANCERS

Mixed review

The number of women selected to be recalled for further assessment, both amongst incidence cancer patients and healthy women, varied significantly; main differences being noted between different screening units, A, B, C and D, (Cochran's Q Test $p < 0.001$) while the two radiologists from the same unit had a similar selection pattern, Table III. The rate of incidence cancers found to be detectable on previous screen images increased with selection rate, as well as the rate of women with incidence cancers selected for the wrong area while specificity for healthy women decreased.

Incidence cancer detection rate varied between 32 % and 15 % and specificity between 69% and 98 %, among the four reviewing units. The unit with the highest detection rate (A) also had the lowest specificity for healthy women as well as the highest proportion of incidence cancer patients selected for the wrong area or opposite breast.

Detection rate (sensitivity) for the in the same (present) round confirmed screen-detected 29 breast cancers was generally high, varying among the units between 86% - 97%, Table III . Units with somewhat lower sensitivity had the highest specificity. Recall rate for lesions confirmed benign in the same (present) round varied widely, between 33 % and 90 %, Table III.

Unit	Radiologist	Women with incidence cancers		Women with screen cancers		Women with benign lesions		Healthy women	
		Correct	Wrong area	Correct	Wrong area	Correct	Wrong area	Selected	Specificity
		No. (%)	No. (%)	No. (%)	No. (%)	No. (%)	No. (%)	No. (%)	%
A	I	28 (24)	14 (12)	25 (86)	0 (0)	17 (81)	1 (5)	187 (24)	76
	II	30 (26)	23 (20)	25 (86)	0 (0)	14 (67)	1 (5)	198 (26)	74
	I+II	37 (32)	23 (20)	27 (93)	0 (0)	17 (81)	1 (5)	242 (31)	69
B	I	20 (17)	1 (1)	25 (86)	0 (0)	7 (33)	1 (5)	26 (3)	97
	II	16 (14)	1 (1)	22 (76)	0 (0)	8 (38)	0 (0)	13 (2)	98
	I+II	17 (15)	1 (1)	25 (86)	0 (0)	7 (33)	0 (0)	18 (2)	98
C	I	34 (29)	13 (11)	28 (97)	1 (3)	15 (71)	2 (10)	96 (12)	88
	II	32 (27)	9 (8)	27 (93)	0 (0)	19 (90)	0 (0)	80 (10)	90
	I+II	35 (30)	9 (8)	28 (97)	1 (3)	19 (90)	0 (0)	85 (11)	89
D	I	23 (20)	6 (5)	24 (83)	0 (0)	13 (62)	0 (0)	57 (7)	93
	II	15 (13)	7 (6)	25 (86)	0 (0)	13 (62)	0 (0)	48 (6)	94
	I+II	23 (20)	5 (4)	25 (86)	0 (0)	13 (62)	0 (0)	60 (8)	92

Table III Mixed review: Women selected correctly and for the wrong area, among 936 reviewed women. Two radiologists single- and double-reading from each of the units A, B, C and D. The mixed set comprised women with incidence cancer (n=117) diagnosed in next screen, with in the present round screen(detected) cancers (n=29) and benign lesions(n=21) and healthy women (n=769). Number of selected and specificity among 769 healthy women.

Non-mixed review

Results from when the 117 incidence cancer patients were reviewed non-mixed (separately), now with reviewers aware that all images represented women with later diagnosed incidence cancer, are shown in Table IV. Detection rate increased significantly (Mc Nemar's test $p = 0.002$) compared to mixed reading, while the differences in detection rate between screening units persisted (Cochran's Q Test $p < 0.001$).

The non-mixed selection- and detection pattern among the screening units was similar to that in mixed review. Selection- and detection rate varied widely: between 36 % and 17 % of women with incidence cancers being correctly selected by the units (Fig 8).

Mixed and non-mixed reviewing

Agreement among the reviewing units as to which women with incidence cancer who should, or should not, be selected by the two methods is shown in figure 9. Ten (9 %) women with incidence cancer were correctly selected by all reviewing units by both methods and constitute the minimum number that could be considered as missed cases due to observer error. On the other hand 62 (53 %) were correctly selected by any of the units using either method and constitute the maximum number that could be regarded as missed cases. This means that 55 (47 %) women were not correctly

selected by any reviewing unit using either method. For most reviewers double-reading increased the number of incidence cancers detected on review compared to single-reading (Table III and IV). The average increase in detection rate was 3.3 (14%) in the mixed and 4.4 (13%) in the non-mixed readings.

Unit	Radiologist	Correct		Wrong area	
		No.	(%)	No.	(%)
A	I	36	(31)	18	(15)
	II	33	(28)	32	(27)
	I+II	42	(36)	32	(27)
B	I	20	(17)	3	(3)
	II	17	(15)	2	(2)
	I+II	20	(17)	2	(2)
C	I	39	(33)	11	(9)
	II	41	(35)	27	(23)
	I+II	48	(41)	20	(17)
D	I	24	(21)	3	(3)
	II	29	(25)	12	(10)
	I+II	27	(23)	5	(4)

Table IV Non-mixed review: Women selected correctly and for the wrong area among incidence cancer patients (n=117). Two radiologists single- and double-reading, from each of the screening units A, B, C and D.

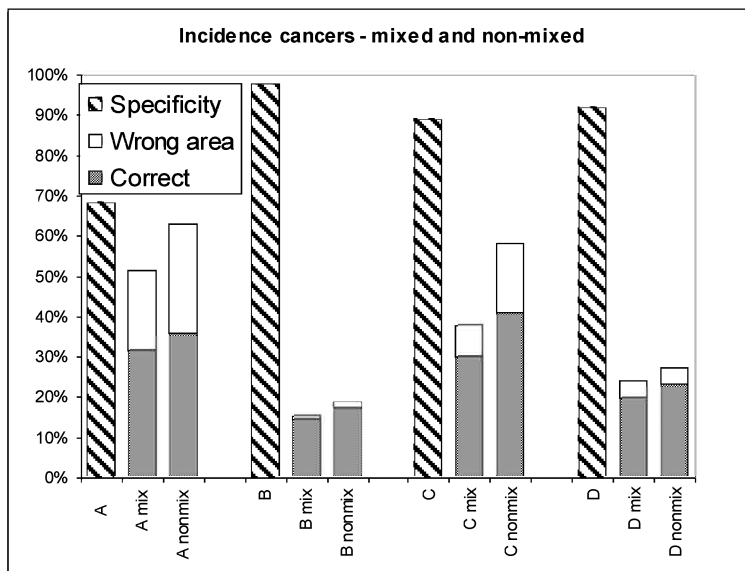


FIG 8 Selection rate (correctly and wrong area selected) of women with incidence cancer (n=117) when reviewed mixed and non-mixed by screening unit A, B, C and D. Specificity among healthy women (n=769).

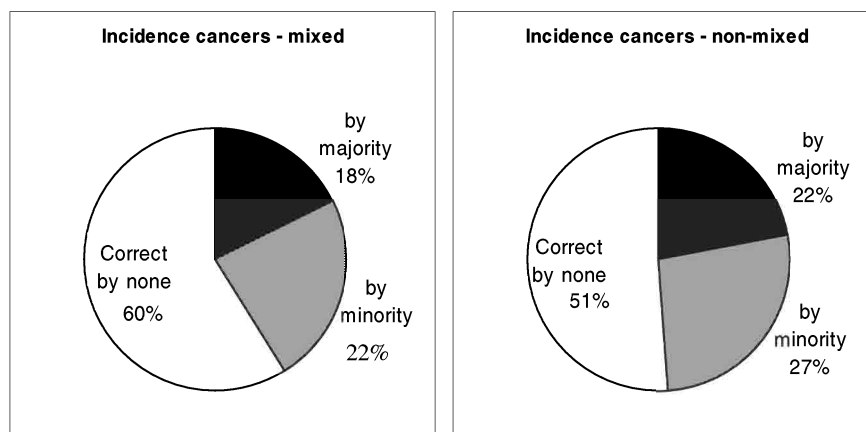


FIG 9 Incidence cancers reviewed mixed and non-mixed (n=117) correctly selected by none, a majority or a minority of the participating screening units.

4.3 PAPER III - COMPUTER AIDED DETECTION OF INTERVAL CANCER

CAD system alone

Review results are summarized in Table V and Fig 10. CAD correctly selected 13 (22%) of the 58 reviewed interval cancer patients. If a woman with a marked vessel calcification, in the same location as the later diagnosed interval cancer, is included altogether 14 (24 %) were detected by CAD. CAD selected 27 (47%) women with interval cancer for the wrong area, 28 (48%) if the woman with marked vessel-calcification is included.

CAD identified all 5 of the in the same round accessible and confirmed screen-detected cancers but only 1 of the included 5 benign lesions.

Specificity for the 211 healthy women was 38%, CAD falsely marking 0.482 signals per image (0.326 false positive spiculations and 0.156 calcifications).

Mixed reviewing - Radiologists without CAD

Selection- and detection rate (correctly selected) for the three reviewers are shown in Table V and fig. 10. Reviewer A had the highest interval cancer detection rate, 17 correctly selected of the 58 interval cancers (29%), reviewer B 12 (21%) and C 11 (19%). Reviewer A selected 14 women with interval cancer for the wrong area or opposite breast (24%) compared to B 5 (9%) and C 6 (10%).

For screen-detected cancers and benign lesions the number of correctly identified lesions was high for all reviewers, Table V. Specificity among the 211 healthy women varied between 73 % and 89 %. As expected the reviewer with the lowest specificity had the highest interval cancer detection rate and vice versa.

Mixed reviewing - Radiologists with CAD

Reviewing the mixed set with aid of CAD made little difference to reviewers' detection rate compared to the reading without CAD, Table V. The maximum change in detection rate was 2 women (3 %). The number of women with interval cancer selected for the wrong area was reduced and specificity among healthy women increased for all

three reviewers (73%, 82% and 89% without CAD compared to 78%, 90% and 92% with CAD) when using CAD, Fig. 10.

Mixed	Radiologist			
	A	B	C	CAD
Interval cancers (n=58)	No. (%)	No. (%)	No. (%)	No. (%)
without CAD, correct	17 (29)	12 (21)	11 (19)	14 (24)
wrong area	14 (24)	5 (9)	6 (10)	27 (47)
with CAD, correct	16 (28)	10 (17)	13 (22)	* (*)
wrong area	9 (16)	1 (2)	5 (9)	* (*)
Screen detected cancers (n=5)				
without CAD, correct	5 (100)	4 (80)	5 (100)	5 (100)
wrong area	0 (0)	0 (0)	0 (0)	0 (0)
with CAD, correct	5 (100)	4 (80)	5 (100)	* (*)
wrong area	0 (0)	0 (0)	0 (0)	* (*)
Benign lesions (n=5)				
without CAD, correct	3 (60)	3 (60)	3 (60)	1 (20)
wrong area	0 (0)	0 (0)	0 (0)	1 (20)
with CAD, correct	3 (60)	3 (60)	3 (60)	* (*)
wrong area	0 (0)	0 (0)	0 (0)	* (*)
Healthy women (n=211)				
specificity without CAD	154 (73)	174 (82)	187 (89)	81 (38)
specificity with CAD	164 (78)	190 (90)	194 (92)	* (*)
Non-mixed				
Interval cancers (n=58)	No. (%)	No. (%)	No. (%)	No. (%)
correct	21 (36)	17 (29)	19 (33)	14 (24)
wrong area	26 (45)	7 (12)	9 (16)	27 (47)

*) not applicable

Table V Women selected correctly and for the wrong area at mixed (280) and non-mixed (58) review by radiologist A, B and C with and without CAD, and for CAD alone. The mixed set comprised interval cancers (58), screen-detected cancers (5), benign lesions (5), and healthy women (211).

Non-mixed reviewing

Interval cancer detection rate increased for all three radiologists when reviewing the interval cancers separately (non-mixed), with reviewers aware that all reviewed images represented women with a subsequent interval cancer, Table V fig. 10. Detection rate varied between 36 % and 29 % for the three reviewers and the increase in detection rate compared to mixed reading without aid of CAD varied between 24 % and 73 %. The number of women with interval cancer selected for the wrong area also increased.

Radiological characteristics at preceding screen

Only 5 of the 58 interval cancers had the appearance of spiculations and/or microcalcifications at the previous screen. All five were correctly selected by CAD while only one of the radiologists identified all five. All interval cancers detected by CAD were also identified by at least one of the radiologists. One of five asymmetries/distortions was detected by CAD while all five were identified by at least

one radiologist. Of the two nodular cancers one was selected by CAD, both by all radiologists. Of three cancers with the appearance of focal masses, not detected by CAD, two were selected by one reviewer in mixed and the third in non-mixed reading. Of the 11 parenchym-like densities CAD correctly selected 6 (7 if including the marked vessel calcification) which is in the same range or better than the radiologists.

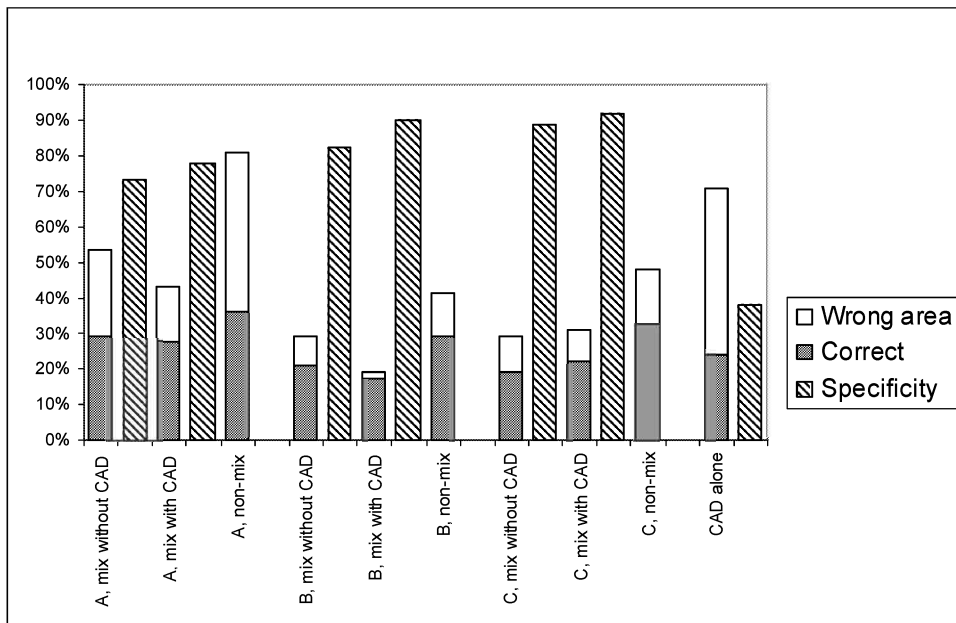


FIG 10 Women with interval cancer selected correctly and for the wrong area by reviewer A, B and C without and with aid of CAD and for the CAD system alone. Specificity among healthy women (n=211).

The potential sensitivity increase for double reading, comparing CAD to the radiologists, summarised for each reviewer was as follows: Reviewer A had the potential opportunity to detect 5 more interval cancers (in mixed review without CAD) if double reading with B, 1 with C and 5 with CAD. For reviewer B corresponding figures are 10 with A, 3 with C and 8 with CAD, and for reviewer C 7 with A, 4 with B and 8 with CAD.

4.4 PAPER IV - FOLLOW-UP OF INCIDENCE CANCERS

Age distribution

Age distribution according to review category for the 117 incidence cancer patients is shown in table VI. Women with correctly selected tumours had a non-significant tendency to be older at diagnosis than those not correctly selected (p=0.070).

Tumour size, nodal status and tumour stage

Information on tumour size, nodal status, and pTNM stage according to review category for all the 122 tumours are shown in Table VI. Review category was as follows: 60 (49%) not correctly selected, 34 (28%) correctly selected by a minority and 28 (23%) correctly selected by a majority of reviewing units. No significant differences in tumour size, nodal status and pTNM stage between tumours of different review

category was noted, even when tumours correctly selected were considered as one group, or when tumours correctly selected in mixed reading were considered separate from those correctly selected in non-mixed review (p=0.458, p=0.281 and p=0.321 respectively).

	Not correct n (%)	Correct by minority n (%)	Correct by majority n (%)	All n (%)
Age at diagnosis				
50-59	22 (40)	9 (26)	4 (14)	35 (30)
60-69	33 (60)	25 (74)	24 (86)	82 (70)
sum	55 (100)	34 (100)	28 (100)	117 (100)
mean age	61.3	62.1	63.9	62.2
Tumour size (mm)				
≤10	16 (32)	12 (40)	12 (46)	40 (38)
11-20	21 (42)	13 (43)	10 (38)	44 (42)
21-30	10 (20)	4 (13)	0 (0)	14 (13)
>30	3 (6)	1 (3)	4 (15)	8 (8)
sum	50 (100)	30 (100)	26 (100)	106 (100)
mean diameter (mm)	15.8	14.2	14.5	15.0
N/A	2		1	3
Nodal status				
negative	36 (71)	26 (84)	21 (78)	83 (76)
positive	15 (29)	5 (16)	6 (22)	26 (24)
1-3	10 (20)	5 (16)	5 (19)	20 (18)
≥4	5 (10)	0 (0)	1 (4)	6 (6)
sum	51 (100)	31 (100)	27 (100)	109 (100)
pTNM Stage				
0	9 (15)	3 (9)	1 (4)	13 (11)
I	27 (45)	21 (62)	17 (61)	65 (53)
II	24 (40)	10 (29)	10 (36)	44 (36)
III+IV	0 (0)	0 (0)	0 (0)	0 (0)
sum	60 (100)	34 (100)	28 (100)	122 (100)
Histology				
Ductal (incl tubular)	41 (68)	22 (65)	20 (71)	83 (68)
lobular	6 (10)	8 (24)	7 (25)	21 (17)
CIS	9 (15)	3 (9)	1 (4)	13 (11)
other	4 (7)	1 (3)	0 (0)	5 (4)
sum	60 (100)	34 (100)	28 (100)	122 (100)
Grade (invasive)				
I	10 (21)	6 (20)	11 (41)	27 (26)
II	27 (56)	20 (67)	13 (48)	60 (57)
III	11 (23)	4 (13)	3 (11)	18 (17)
sum	48 (100)	30 (100)	27 (100)	105 (100)
N/A	4			4

Table VI Tumour characteristics according to review category in incidence cancer patients (n=117), five with bilateral malignancies,

Two women (one with bilateral malignancies) who received neoadjuvant chemotherapy are included in the table with their TNM stage prior to neoadjuvant treatment but excluded from (postoperative) tumour size analysis (N/A=not applicable). Mean tumour diameter was 15 mm with 26 (21 % of all tumours, 24 % of the invasive tumours) node positive. A total of 78 (64 %) tumours were at stage \leq I and 44 (36 %) at stage \geq II. Out of the 62 correctly selected tumours 42 (68%) were at stage \leq I and 20 (32%) at stage \geq II.

Histopathology and tumour grade

Histopathology of the 122 tumours, summarized in Table VI, showed 3 tubular cancers (2.5 %), 37 (30.3 %) tubuloductal, 43 (35.2%) ductal, 21 lobular (17.9 %), 5 of other type (4.1 %) (including medullary, cystosarcoma phylloides, papillary, mucinous and metaplastic cancer) and 13 (10.7 %) pure cancer in situ without any infiltrating focus. There were no significant differences according to review category but a slight tendency to find more numerous lobular cancers and fewer cases of cancer in situ among the retrospectively identified cases was noted ($p = 0.130$). Histological tumour grade could be established for 105 of the 109 invasive tumours, Table VI. Incidence cancers correctly selected by a majority of reviewers had a slight non-significant tendency to be of lower grade than had tumours not correctly selected ($p = 0.237$).

Tumour stage according to reviewing unit results - mixed review

For each participating screening unit, A, B, C and D, tumour stage (stage \leq I or \geq II) of both the correctly and the not correctly selected incidence cancer patients is shown in Fig 11 and Table VII (none of the five women with bilateral malignancies was correctly selected for both breasts).

Stage distribution among the 47 correctly selected: 31/47 (66%) at stage \leq I and 16/47 (34%) at stage \geq II, did not differ much from stage distribution among those not correctly selected which was 43/70 (61%) at stage \leq I and 27/70 (39%) at stage \geq II ($p=0.697$).

Comparison between the unit with the highest selection rate (A) and the one with the lowest (B) shows that unit A detected four additional advanced cancers (pTNM stage \geq II) when compared to unit B, representing those women who probably would have benefited the most from an earlier detection of their cancer. Unit A also had the highest rate of selected healthy women, 242 whereas unit B had 18 (Table III). The gain in sensitivity for retrospectively detected tumours stage \geq II at diagnosis by unit A and C compared to B and D was 4 women. The loss in specificity for unit A was constituted by 224 more healthy women being selected by A than by B (242-18) and 182 more healthy women selected than by D, while corresponding figures for unit C were 67 and 25 respectively.

Tumour stage according to reviewing unit results - nonmixed review

In the non-mixed readings detection rate increased for all units, for three units it also increased when including stage \geq II tumours (Table VII). Stage distribution among the 58 women correctly selected in the non-mixed reading was 39/58 (67%) stage \leq I and 19/58 (33%) stage \geq II, and was not significantly different from stage distribution among those not correctly selected which was 39/64 (61%) stage \leq I and 25/64 (39%) stage \geq II.

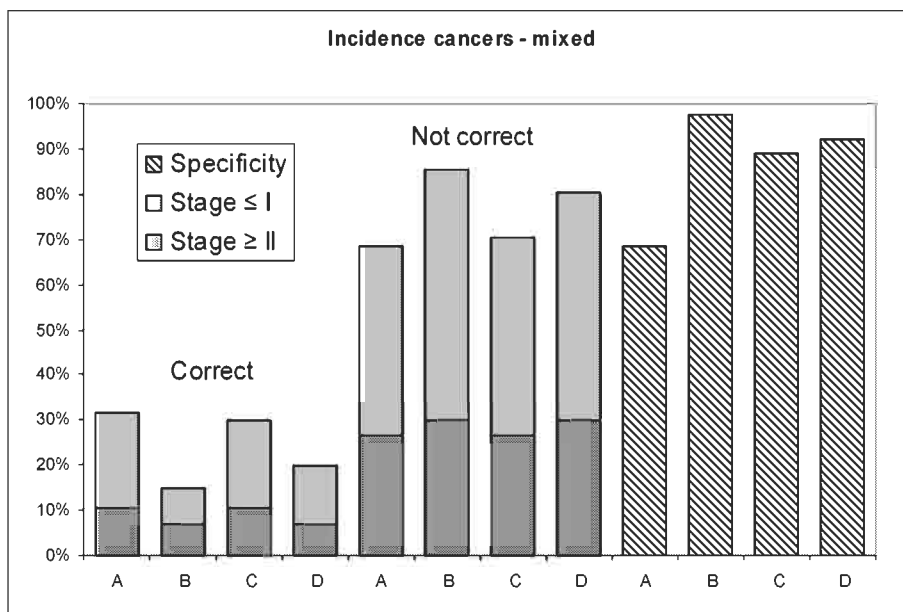


FIG 11 Tumour stage at presentation ($pTNM \leq I$ or $\geq II$) for women with incidence cancer correctly and not correctly selected ($n = 117$) reviewed mixed by reviewers from units A, B, C and D. Specificity among healthy women ($n = 769$).

Stage	Mixed								Non-Mixed							
	Correct				Not correct				Correct				Not correct			
	$\leq I$		$\geq II$		$\leq I$		$\geq II$		$\leq I$		$\geq II$		$\leq I$		$\geq II$	
	No	%	No	%	No	%	No	%	No	%	No	%	No	%	No	%
A	25	68	12	32	49	61	31	39	26	62	16	38	48	64	27	36
B	9	53	8	47	65	65	35	35	13	65	7	35	61	63	36	37
C	23	66	12	34	51	62	31	38	33	69	15	31	41	59	28	41
D	15	65	8	35	59	63	35	37	17	63	10	37	57	63	33	37

Table VII Tumour stage at presentation ($pTNM \leq I$ or $\geq II$) for women with incidence cancer correctly and not correctly selected ($n = 117$) reviewed mixed and non-mixed by reviewers from screening units A, B, C and D.

Clinical follow-up according to review category

During the observation time, median follow-up 8.25 years, ten incidence cancer patients experienced a distant recurrence of the disease. Five of these ten women had incidence cancers not retrospectively identified while five had tumours that were correctly selected by the reviewers: two by a majority and three by a minority, Fig. 12. Five of the ten women with distant recurrences had died from the disease during the follow-up period: in two of these five women an abnormality was identified in retrospect by a majority of reviewing units, in the remaining three women no lesion was identified. Neither were any abnormalities found in the seven women with local recurrences. Two of the original 117 women could not be followed up for presence of recurrence for the last 1.5 and 2.5 years of the observation period respectively. Both

women had stage I tumours and were alive at the end of the observation period, one was not correctly selected while the other was correctly selected by a majority.

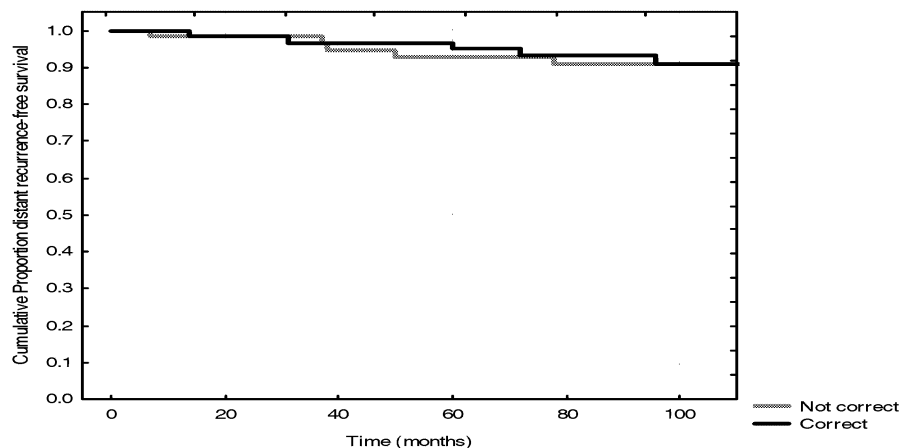


Fig 12 Life-table curves showing distant recurrence-free survival since time of diagnosis (months) for 117 women with incidence breast cancer, correctly ($n = 62$) and not correctly ($n = 55$) selected women ($p=0,898$).

Parenchymal pattern, tumour location and radiological appearance

Detailed information as to parenchymal pattern and tumour location is presented in Table VIII. Overall breast density was very similar between correctly and not correctly selected tumours ($p = 0.809$), while the correctly identified cases had a tendency to more often have DY and less often N1 tumour area pattern ($p = 0.058$).

There were no significant differences in tumour location according to review category ($p = 0.585$), this was also the case when considering the retroglandular location separate ($p = 0.851$). Forty-five (37%) of the 121 mammographically localizable malignant lesions were located in the retroglandular area (Fig 3), 6 (5%) in the retroareolar region and 4 (3%) in the medial half of the breast (12% of women with two-view screens).

The radiological appearance of the correctly identified cases at the preceding screen ($n=62$), in figure 11a shown according to review category, showed calcifications only in 4 (6%), spiculated lesions (with or without calcifications) in 8 (13%), asymmetric density in 5 (8%), distortion in 13 (21%), circumscribed mass (with or without calcifications) in 8 (13%) and unspecific density (with or without calcifications) in 20 (32%). Four (6%) cases were unclassifiable, all selected by a minority in the non-mixed reading.

Corresponding figures at diagnosis ($n = 122$), Fig. 13 b, showed: calcifications only 14 (12%), spiculated lesions (with or without calcifications) 61(50 %), asymmetric density 0 (0 %), distortion 4 (3 %), circumscribed mass (with or without calcifications) 36 (29%) and unspecific density (with or without calcifications) 4 (3 %). In 3 (2.5%) cases the lesion was mammografically occult at diagnosis.

Tumours correctly selected were more likely to present as spiculated masses and less likely to present as circumscribed lesions and calcifications compared to incidence

cancers not correctly selected. Although the differences reached significance ($p=0.003$) there was, as seen in figure 13, a substantial overlap between all categories of radiological appearance among all review categories. There were no significant differences in radiological appearance according to tumour stage (paper IV).

Recall rate, false assessments and false negative tumours

Six of the 117 (5%) women had been recalled in the preceding screen, a higher rate than the ordinary recall rate at that time (approximately 1.5 %). In three of these six women, all three correctly selected by a majority, the recall assessment concerned the lesions later diagnosed as incidence cancers. Another seven women had been recalled in the round prior to the immediately preceding screen, five of them because of the lesions later diagnosed as incidence cancers, all five correctly selected: three by a majority and two by a minority of reviewing units. To sum up: in eight of the 117 incidence cancer patients, all eight correctly selected by the reviewers, the lesion had previously been falsely assessed as benign.

Twenty-eight tumours were considered false negative on the basis of being correctly selected by a majority of reviewing units. At least 14 of those 28 had probably been misinterpreted rather than overlooked; 6 were among the recalled falsely assessed cases (see above) and 8 lesions were visible in images prior to the immediately preceding screen. Ten of the 28 (36%) false negative tumours were at stage \geq II by the time of diagnosis, this being the same proportion as for all 122 incidence cancers. Among the 14 lesions regarded as misinterpreted 3 (21%) were at stage \geq II at presentation.

	Not correct n (%)	Correct n (%)	All n (%)
Parenchymal pattern			
N1	11 (18)	15 (24)	26 (21)
P1	34 (57)	30 (48)	64 (52)
P2	14 (23)	16 (26)	30 (25)
Dy	1 (2)	1 (2)	2 (2)
sum	60 (100)	62 (100)	122 (100)
Parenchymal pattern tumour area			
N1	25 (42)	21 (34)	46 (38)
P1	18 (31)	21 (34)	39 (32)
P2	16 (27)	15 (24)	31 (26)
Dy	0 (0)	5 (8)	5 (4)
sum	59 (100)	62 (100)	121 (100)
N/A	1		1
Tumour location			
In retroglandular area	21 (36)	24 (39)	45 (37)
Other locations	38 (64)	38 (61)	76 (63)
sum	59 (100)	62 (100)	121 (100)
N/A	1		1

Table VIII Parenchymal pattern and tumour location of incidence cancers (n=122) according to review status.

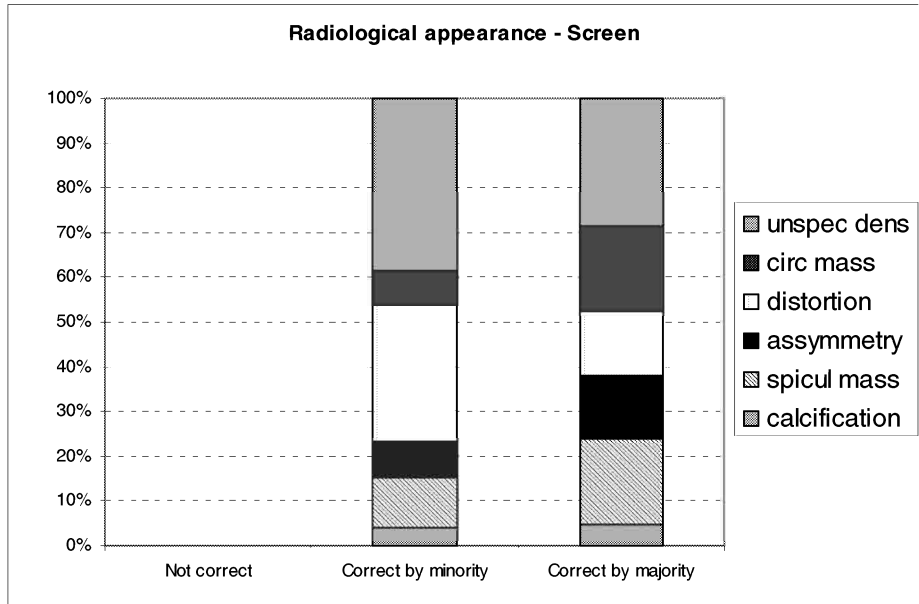


Fig 13a

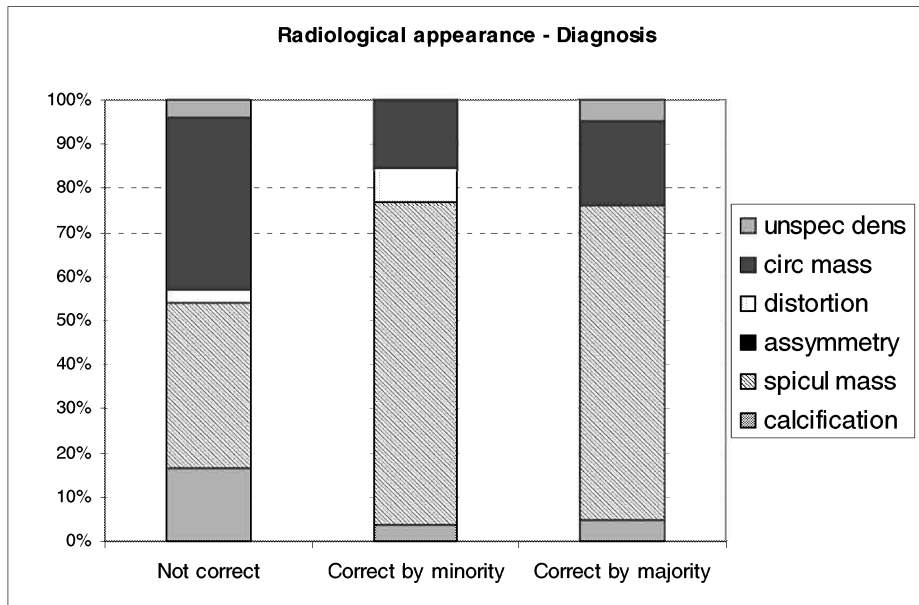


Fig 13b Radiological appearance at preceding and presenting screen for incidence cancers according to review category ("Not correct" "Correct by minority" "Correct by majority").

4.5 PAPER V - FOLLOW-UP OF INTERVAL CANCERS

Age distribution and time interval

Women with interval tumours correctly selected by a majority of reviewing units tended to be older ($p=0.124$) with shorter mean time interval between screen and diagnosis (12.5 months) compared to patients with interval cancers not correctly selected (15.6 months), ($p=0.088$), Table IX.

Tumour size, nodal status and tumour stage

Of all the 103 malignancies 66 (64%) were not correctly selected, 15 (15%) were correctly selected by a minority and 22 (21%) were correctly selected by a majority. Information on tumour size, nodal status, and pTNM stage according to review category are shown in Table IX. There were no significant differences between tumours of differing review category, even when correctly selected tumours were considered as one group, or when tumours correctly selected by mixed were considered separate from those detected non-mixed (p -values between 0.359 and 0.965). The two not surgically treated women and the four women who received neoadjuvant chemotherapy are included in the table with their TNM stage prior to neoadjuvant treatment but are excluded from (postoperative) tumour size analysis (N/A=not applicable). Mean tumour diameter was 18.1 mm with 59 (66%) node negative and 31 (34 %) node positive tumours. A total of 57 (55 %) tumours were stage \leq I and 46 (45 %) stage \geq II.

Histopathology and tumour grade

Histopathology of the 101 surgically treated malignancies is summarized in Table IX. Three were tubular cancers, 14 tubuloductal, 49 ductal, included in the ductal category in table IX. Sixteen malignancies were lobular, 10 of other type (including medullary, mucinous, papillary, adenoid cystic) and 9 pure cancers in situ (CIS) without any infiltrating focus.

Histological tumour grade could be established for 83 of the 92 surgically removed invasive tumours (Table IX). There were no significant differences in the distribution of histological tumour type or tumour grade between tumours of differing review category ($p=0.835$ and 0.611 respectively).

Tumour stage according to reviewing unit results -mixed review

Tumour stage (\leq I or \geq II) for both correctly and not correctly selected interval patients is shown in Table X and figure 14. Although specificity for healthy women varied substantially between the reviewing units, with 89% for unit A, 97% for B and 92% for unit C, the number of correctly identified tumours, including tumours at stage \geq II at presentation, was at almost the same level for all units. Stage distribution among the 13 women with identified tumours, 7/13 (54%) at stage \leq I and 6/13 (46%) at stage \geq II, was almost identical to stage distribution among those not identified, with 25/45 (55%) at stage \leq I and 20/45 (44%) at stage \geq II ($p=1.0$).

Non-mixed review

Interval cancer detection rate increased in the non-mixed reading compared to mixed, including tumours stage \geq II at presentation, Table X. Stage distribution among the 37 women with tumours correctly selected, 22/37 (59%) at stage \leq I and 15/37 (41%) at

	Not correct n (%)	Correct by minority n (%)	Correct by majority n (%)	All n (%)
Age at diagnosis				
50-59	36 (55)	7 (47)	6 (27)	49 (48)
60-69	21 (32)	6 (40)	11 (50)	38 (37)
≥70	8 (12)	2 (13)	5 (23)	15 (15)
sum	65 (100)	15 (100)	22 (100)	102 (100)
mean age	60.5	60.2	63.3	61.0
Interval screen - diagnosis				
≤1 year	20 (31)	2 (13)	11 (50)	33 (32)
> 1 year	45 (69)	13 (87)	11 (50)	69 (68)
sum	65 (100)	15 (100)	22 (100)	102 (100)
mean interval (months)	15.6	14.9	12.5	14.8
Tumour size (mm) of invasive				
≤10	12 (21)	3 (25)	5 (28)	20 (23)
11-20	26 (45)	6 (50)	7 (39)	39 (44)
21-30	16 (28)	2 (17)	4 (22)	22 (25)
>30	4 (7)	1 (8)	2 (11)	7 (8)
sum	58 (100)	12 (100)	18 (100)	88 (100)
mean diameter (mm)	18.4	16.8	17.1	18.1
N/A	3	2	1	6
Nodal status				
negative	35 (61)	10 (71)	14 (74)	59 (66)
positive	22 (39)	4 (29)	5 (26)	31 (34)
1-3	14 (25)	3 (21)	2 (11)	19 (21)
≥4	8 (14)	1 (7)	3 (16)	12 (13)
sum	57 (100)	14 (100)	19 (100)	90 (100)
N/A	4			4
pTNM Stage				
0	5 (8)	1 (7)	3 (14)	9 (9)
I	30 (45)	9 (60)	9 (41)	48 (47)
II	27 (41)	3 (20)	9 (41)	39 (38)
III+IV	4 (6)	2 (13)	1 (5)	7 (7)
sum	66 (100)	15 (100)	22 (100)	103 (100)
Histology				
Ductal (incl tubular)	44 (69)	10 (67)	12 (55)	66 (65)
Lobular	8 (13)	3 (20)	5 (23)	16 (16)
CIS	5 (8)	1 (7)	3 (14)	9 (9)
Other	7 (11)	1 (7)	2 (9)	10 (10)
sum	64 (100)	15 (100)	22 (100)	101 (100)
N/A	2			2
Grade (invasive)				
I	15 (28)	2 (15)	3 (18)	20 (24)
II	16 (30)	6 (46)	8 (47)	30 (36)
III	22 (42)	5 (38)	6 (35)	33 (40)
sum	53 (100)	13 (100)	17 (100)	83 (100)
N/A	7	1	3	11

Table IX (on previous page) Tumour characteristics of interval cancers (n=103) according to review category.

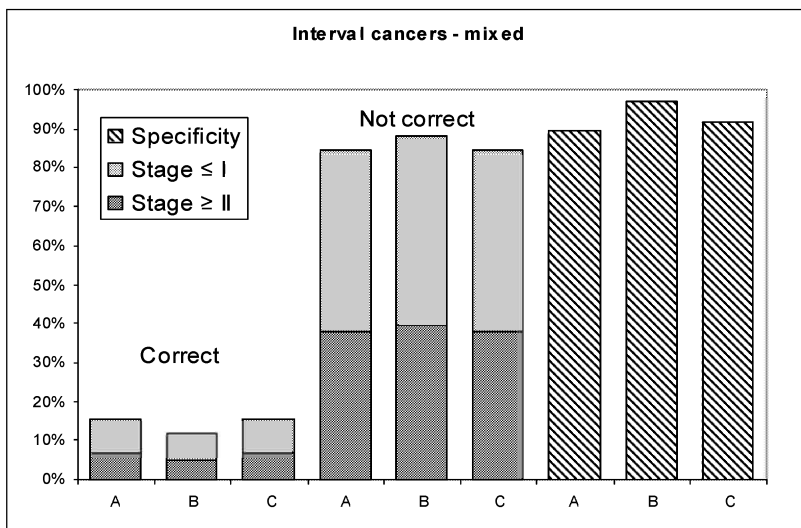


Fig 14 Tumour stage at presentation (pTNM ≤ I or ≥ II) of correctly and not correctly selected women with interval cancer (n = 58) reviewed mixed by reviewers from unit A, B, and C. Specificity among healthy women (n = 390).

stage ≥ II, did not differ much from the total of 65 not correctly selected, 43/65 (52%) at stage ≤ I and 31/65 (48%) at stage ≥ II (p=0.539).

Stage	Mixed								Non-Mixed							
	Correct				Not correct				Correct				Not correct			
	≤ I		≥ II		≤ I		≥ II		≤ I		≥ II		≤ I		≥ II	
No	%	No	%	No	%	No	%	No	%	No	%	No	%	No	%	
A	5	56	4	44	27	55	22	45	20	65	11	35	36	51	35	49
B	4	57	3	43	28	55	23	45	15	60	10	40	41	53	36	47
C	5	56	4	44	27	55	22	45	14	52	13	48	42	56	33	44
D	-	-	-	-	-	-	-	-	8	44	10	56	48	57	36	43

Table X Tumour stage at presentation (pTNM ≤ I or ≥ II) for women with interval cancer correctly and not correctly selected, reviewed mixed (n = 58) and non-mixed (n=102) by reviewers from screening units A, B, C and D.

Clinical follow-up related to review result

Thirteen out of the 101 interval cancer patients included in survival analysis (excluding two women with a previous invasive breast cancer in the opposite breast and including one woman not radiologically reviewed due to missing screen images) had died from breast cancer during the observation period, with a median follow-up 107 months (Fig. 15). Another three women suffered a distant recurrence of the disease while three patients had a local recurrence, all six still being alive at the end of the follow-up period.

Ten out of the 13 women who died from breast cancer had tumours that were not correctly selected by the reviewers. Two had tumours identified in retrospect (one by majority and one by minority) and one woman was not included in the radiological review. Of the three surviving women with distant recurrence one had a tumour which was not correctly selected while two were correctly selected by a minority in retrospect. Among the three patients with only local recurrences two had cancers not correctly selected while the third was correctly selected by a minority.

There were no significant differences in breast cancer-specific survival and distant recurrence-free survival (Fig. 16) between women with interval cancers retrospectively identified compared to those not identified ($p=0.106$ and 0.322 respectively).

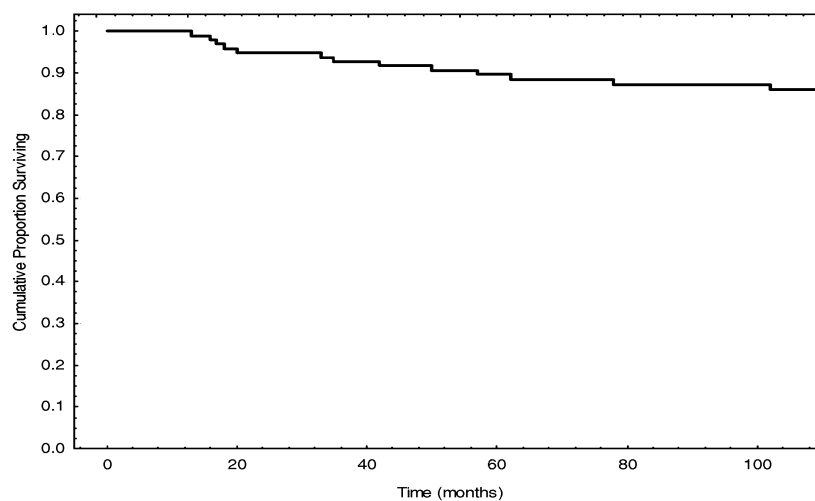


Fig 15 Life-table curve showing breast cancer specific survival for 101 women with interval breast cancer, time from diagnosis (months).

Parenchymal pattern, tumour location and appearance

No significant differences in breast density, both overall and in tumour area, were observed when comparing tumours correctly selected to those not correctly selected, Table XI ($p=0.911$ and 0.470 respectively). There were no significant differences in tumour location (Table XI) between tumours correctly selected and those not correctly selected ($p=0.847$), nor when only tumours located in the retroglandular area (Fig. 3) were considered ($p=0.954$). Of the 87 tumours that were possible to localize exactly on the mammogram, 36 (41%) were located in the retroglandular area. Another 5 (6%) lesions were in the retroareolar region and 14 located in the medial half of the breast (20% of the 69 with two-view examinations).

There was no significant difference in radiological tumour appearance according to review category (Fig.17 a and b) or in tumour stage (paper V) at diagnosis for women with radiologically classifiable findings at preceding screen or at diagnosis. Radiological appearance at the preceding screen for correctly selected cases ($n=37$) showed calcifications only in 7 (19%), spiculated lesion in 1 (3%), asymmetric

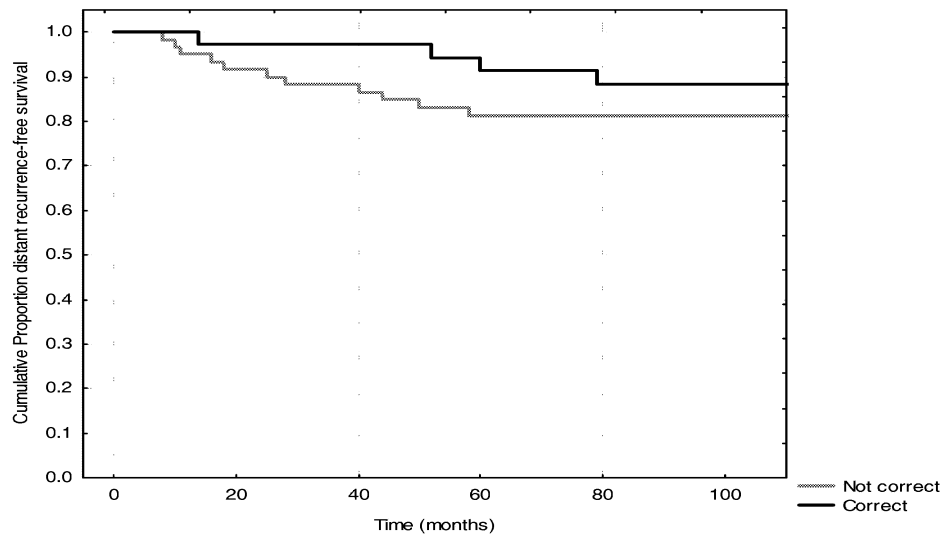


Fig 16 Life-table curves showing distant recurrence-free survival for women with interval breast cancer correctly ($n = 37$) and not correctly selected ($n = 63$), time from diagnosis (months).

	Not correct n (%)	Correct n (%)	All n (%)
Parenchymal pattern			
N1	9 (14)	4 (11)	13 (13)
P1	33 (50)	21 (57)	54 (52)
P2	17 (26)	9 (24)	26 (25)
Dy	7 (11)	3 (8)	10 (10)
sum	66 (100)	37 (100)	103 (100)
Parenchymal pattern tumour area			
N1	20 (40)	9 (24)	29 (33)
P1	14 (28)	14 (38)	28 (32)
P2	11 (22)	9 (24)	20 (23)
Dy	5 (10)	5 (14)	10 (11)
sum	50 (100)	37 (100)	87 (100)
N/A	16		16
Tumour location			
In retroglandular area	20 (40)	16 (43)	36 (41)
Other locations	30 (60)	21 (57)	51 (59)
sum	50 (100)	37 (100)	87 (100)
N/A	16		16

Table XI Parenchymal pattern and tumour location of interval cancers ($n=103$) according to review status.

density in 1 (3%), distortion in 8 (22%), circumscribed mass in 5 (14%) and unspecific density in 13 (35%). Two (5%) cases, selected by a minority, were unclassifiable. Corresponding figures for women with tumours manifested as a pathological mammographic finding by the time of diagnosis ($n = 86$) showed: calcifications only 11 (13%), spiculated lesions 34 (40%), asymmetric density 7 (8%), distortion 6 (7%),

circumscribed mass 20 (23%), unspecific density 8 (9 %). At diagnosis fourteen tumours were mammographically occult and for three women images from diagnosis were missing, none of these three selected by any unit.

Recall rate, false assessments and false negative tumours

Four of the 102 (3.9%) reviewed women had been recalled in the preceding screen, a somewhat higher number than the ordinary recall rate at that time (approximately 1.5 % at unit D responsible for 57% of the reviewed examinations and 3.8 % at unit C). In all four cases the recall assessment concerned the lesion later diagnosed as an interval cancer (one of the cases referred because of a clinical finding without any mammographic abnormality). Two more patients from unit D had been recalled at the screen occasion prior to the immediately preceding screen, one of them because of the lesion later diagnosed as an interval cancer.

The 22 interval cancers identified by a majority of reviewing units are considered to be false negatives. Four of these had been recalled and falsely assessed negative (see above) and in another four cases from unit D the lesions were clearly visible on preceding screen images and almost unchanged when compared to older images.

To sum up: in at least eight of the 22 false negative tumours, six from unit D and two from unit C, we regard the malignancies as having been misinterpreted rather than overlooked on the basis of them being either falsely assessed or clearly visible in mammograms prior to the immediately preceding screen.

Of the 22 false negative malignancies 12 (55%) were at stage \leq I and 10 (45%) were at stage \geq II, the same stage distribution as for the whole group of interval cancer patients. Of the 8 tumours regarded as misinterpreted 6 (75%) were at stage \leq I and 2 (25%) at stage \geq II.

4.6 COMPARISON BETWEEN INCIDENCE AND INTERVAL CANCERS

The rate of false negative malignancies, defined as tumours correctly selected by a majority of reviewing units, were almost the same in the two cancer patient groups while the rate not correctly selected was higher among interval cancer patients. The reviewers' incidence cancer detection rate more closely followed selection rate, resulting in substantial differences in reviewers' sensitivity, while interval cancer detection rate was very similar among the reviewers and less affected by selection rate. The interval cancer patients tended to be younger at diagnosis compared to patients with incidence cancer, with mean ages being 61.0 and 62.2 years respectively ($p=0.144$). The interval tumours were significantly larger than the incidence cancers, with a mean diameter of 18.1 mm and 15.0 mm respectively, ($p=0.027$), with significantly more cancers at an advanced stage ($p=0.041$), of grade III ($p=0.001$) and tended to more often be node positive ($p=0.115$) when compared to incidence cancers.

There were significantly fewer women with N1 and more women with DY parenchymal breast pattern among interval cancer patients than among women with incidence cancer ($p=0.025$). There were no significant differences in tumour area density, tumour location or histology between the two groups of patients.

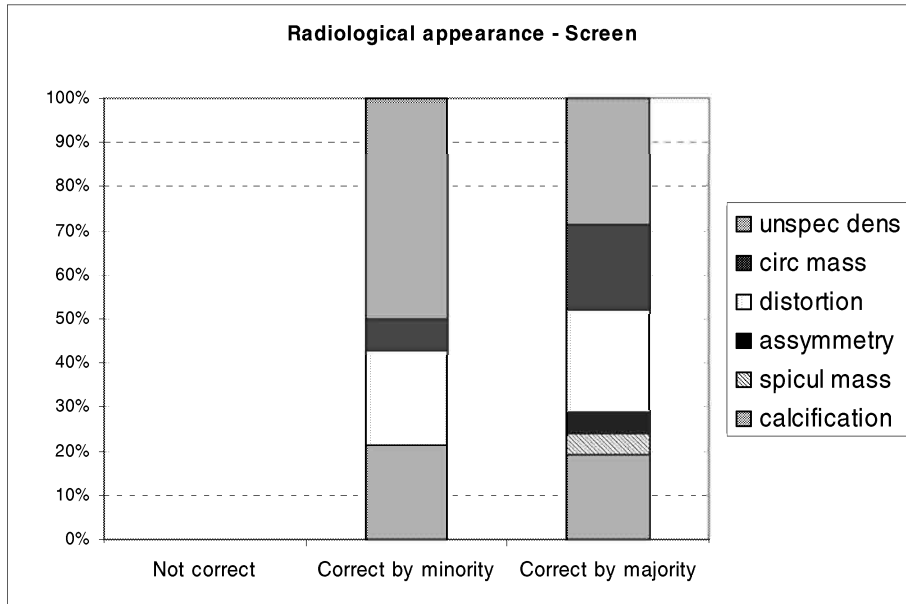


Fig 17a

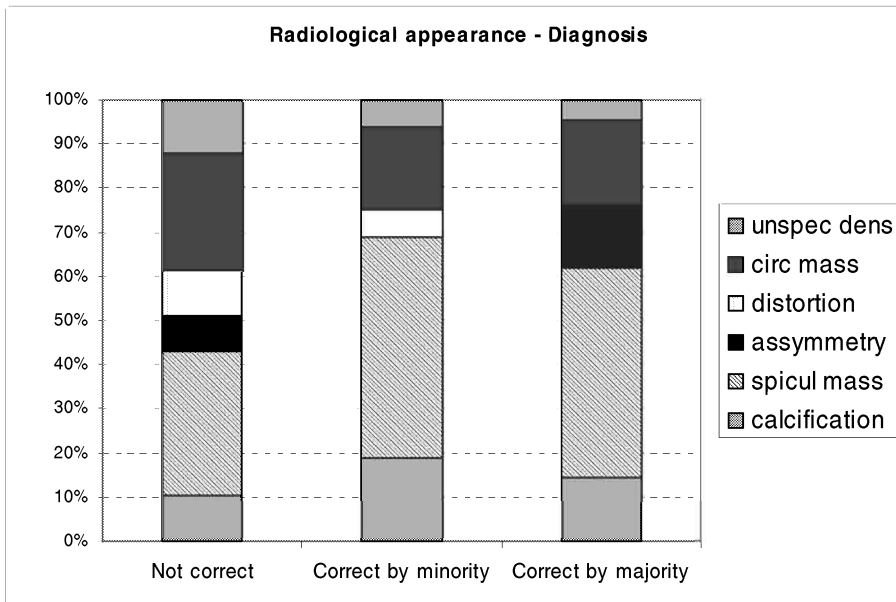


Fig 17b Radiological appearance at preceding screen and at presentation for interval cancers according to review category ("Not correct" "Correct by minority" "Correct by majority").

5 DISCUSSION

5.1 INTERVAL CANCERS (*I*, *V*)

The incidence rate of interval cancers is sometimes considered a quality measurement of screening, especially of radiologists' performance [32][129]. However, comparing *total* interval cancer rates from different studies/screening programmes involves difficulties, as a number of factors apart from the interpreting skill of the radiologists strongly influence these (see 2:2). Furthermore, as the majority of interval cancers in this and other studies are not considered to be false negative, an impairment in image interpretation causing an increased number of false negative cases, must be pronounced before being noticed as an increase in the (total) number of interval cancers. Hence, as a measurement of radiologist' interpretational skills *total* interval cancer rate is an insensitive and inappropriate quality parameter. The rate in this programme, 1.9 per 1000 women screened, was within the range reported in other studies [82-97].

Comparing the rate of interval cancers classified as *false negative* is also problematic, their number being influenced by various factors (see 2:3). This thesis shows that the method used for classification has a substantial impact on the number of interval cancers judged to be false negative. The proportion of interval cancers that possibly could be considered to be missed cases in this review varied between 7% and 34%. The rate depends both on the review method used (mixed or non-mixed) and on the number of reviewers/reviewing units (one of them, a majority or all) deemed necessary to classify a case as missed. Regarding interval cancers correctly selected by a majority of reviewing units as false negative cases, the false negative rate will be 22%, which is in accordance with those in other studies [82-97]. Another 12% were correctly selected by a minority and may correspond to the review category "minimal sign present". A mixed review method, where a few cases of malignancies are randomly mixed with a majority of normal screening mammograms, resulted in significantly fewer identified malignancies than a non-mixed method. Considering the small differences in detection rate noted between the reviewers from the same screening unit in study II, the result is assumed to only very partially be explained by the fact that the mixed review set was single-read while the non-mixed review was double-read by each unit.

The precision of the selection procedure varied. The selection rate of healthy women and of women with interval cancer selected for the wrong area differed substantially among the units whereas interval cancer detection rate was very much the same. This similarity in detection rate was also evident for the subgroup of malignancies at stage \geq II at presentation. Hence, trying to increase sensitivity by lowering the threshold for recall, was in this retrospective study set up found likely to result in too many false positive examinations without a corresponding gain in sensitivity, including sensitivity for stage \geq II tumours.

We could not demonstrate reviewers' selection- and detection patterns to be influenced by whether they represented the screening unit responsible for the original interpretation of the images (internal reviewers) or not (external reviewers). However the result might have been different had the internal reviewers not been aware of the

external reviewers and if less time had passed since the diagnosis, making the reviewers less likely to forget individual cases.

In accordance with other reports [117,118] we found a tendency for shorter mean time interval between screen and diagnosis in false negative tumours compared to true interval cancers. Breast cancer specific survival of 86% for the whole group of women with interval cancer, is almost the same as the survival rate reported by Frisell et al [88] for women with interval cancers within the randomized trial at Stockholm Söder Hospital, diagnosed a few years earlier than the women in this study. Survival rate is also close to that reported in other studies [83, 84]. Mean tumour diameter of 18.1 mm, was smaller than that reported by others (in the Frisell study 21 mm) but was influenced by the fact that a few women with large breast tumour masses were excluded from tumour size analyses due to preoperative neoadjuvant treatment.

Our results did not demonstrate any significant differences in tumour size, nodal status, tumour stage, histological subtype or tumour grade between malignancies correctly selected by the reviewers compared to those not correctly selected. Thus women with an interval cancer correctly selected, with a potentially delayed diagnosis, were not more often found to have an advanced disease at presentation than were interval patients with no detected abnormalities. Accordingly we found no tendency for patients with lesions correctly selected by the reviewers to have an outcome at follow-up which was worse than for those with no detected findings. While this comparison has to be interpreted with caution due to the small numbers involved it is in accordance with other reports on the follow-up of women with overlooked or misinterpreted interval tumours [84][88][93] even though the review procedure in these reports was not blinded, not mixed and not performed by several radiologists, including external, radiologists, as in this study.

No significant differences in parenchymal density, overall or in the tumour area, tumour location or radiological appearance could be demonstrated between malignancies correctly selected by the reviewers and those not identified. Varying results have been reported concerning parenchymal pattern in false negative tumours [70][122] but our findings do not agree with the observation that false negative cancers more likely occur outside the breast parenchymal disc [100]. However the results further emphasize the need for radiologists involved in screen reading to pay continuous attention to abnormalities arising in the retroglandular tissue, the retroareolar region and the medial half of the breast as a substantial proportion of breast cancers occur in these areas [20-21]. No specific radiological feature could be demonstrated as being representative of tumours possible to identify in retrospect or at stage \geq II at presentation which is in accordance with previous studies [84][118]. Not surprisingly and consistent with the natural history of the disease the number of lesions with the appearance of spiculated lesions and circumscribed masses increased while unspecific densities and distortions decreased during the interval between screening and diagnosis.

As reported by others, recall rate at the preceding screen tended to be higher for women with interval cancers, especially among false negatives, than in the general screened population [118][128]. Eight of the 22 false negative tumours might be considered as

misinterpreted rather than overlooked, if summarising cases with lesions clearly visible in images prior to the immediately preceding screen and cases recalled and falsely assessed as benign. Previous studies have shown that tumours with mammographically slow progression can escape detection due to absence of interval change [70]. The proportion of tumours misinterpreted would probably have been higher, had the preceding round at unit C not been the prevalent round with no prior screen images available.

5.2 COMPUTER AIDED DETECTION (CAD) SYSTEM (III)

The sensitivity of the Computer Aided Detection (CAD) system was found to be satisfactory, equal to that of the radiologists and in accordance with previous studies [111,112]. This is somewhat surprising as the preceding screen images of interval cancers often lack the malignant characteristics (spiculations and calcifications) for which CAD is primarily adjusted. All lesions with such characteristics were consequently identified correctly by CAD while only one of the three radiologists detected all of them.

For lesions of type distortions and asymmetries the sensitivity of the CAD system was inferior to that of the radiologists. Detection of such lesions is often based on comparison of images, for instance of the right and the left breast, which the CAD system was not yet able to accomplish. Detection rate for parenchym-like faint densities, which lack typical malignant characteristics, was to our surprise as high for CAD as for the radiologists. In fact, following the high sensitivity of CAD, only one of the three radiologists would have had a potentially higher increase in detection rate if double reading with a colleague instead of double reading with CAD.

Despite this high sensitivity, using CAD had very little influence on the radiologists' detection rate, which gives rise to some reflections. First, cancers missed due to observer error may either have been misinterpreted or overlooked. Observed but misinterpreted cancers can not be prevented by CAD while the substantial number of false negative cases caused by observational oversight [77][94] possibly can. However, in the present review situation, the risk of cancers being overlooked was minimal, all reviewers being focused on their performance. This demonstrates a shortcoming: the complex and partially unknown circumstances causing cancers to be overlooked are difficult to imitate in a study reading situation. Secondly the outcome might have been different if less experienced radiologists had participated in the study, as an earlier study have shown these radiologists to receive most benefit from CAD, [110].

Specificity of the CAD system itself was low (38%), although increased when compared to previous studies [110][112]. In agreement with those studies, no reduction in radiologists' specificity occurred when using CAD. This is important as an increased number of healthy women involved in further medical assessment would mean a serious objection to the use of CAD.

However the many false positive markings have probably reduced the radiologists' confidence in CAD, making it less likely that CAD would lead them to revise their

original interpretations (indicated by high radiologist specificity being unaffected by the use of CAD). If radiologists would revise their original interpretations according to CAD, the potential benefit of CAD as a double-reader may be realized. Therefore improving specificity, preferably with preserved sensitivity, will probably make CAD more useful as a double-reader. It should also reduce image reading time since fewer false positives would reduce the time spent by the radiologist in evaluating them, which is of importance to radiologists reading large numbers of screening mammograms. Possible legal aspects, following CAD markings judged to be false positives by the radiologist and therefore ignored, could also be a matter of concern.

In non-mixed review the increase in detection rate was less pronounced than in the previous study (I) but more marked than in a study by Duncan and Wallis [78]. Different reviewers and differences in mixing proportions (1:8 in the earlier study, 1:3 in the study by Duncan and Wallis and 1:5 in the present study) are the probable causes of this.

No matter what improvements in sensitivity/specificity will be achieved, CAD could never fully replace a double-reading radiologist as he/she is not limited to acting purely as a perceptual aid, but also takes part in the further discussions necessary when examinations are difficult to interpret.

To demonstrate the final usefulness large scale prospective studies of radiologists using CAD in every day mass mammography screening practice is necessary, preferably after adjusting the CAD programme to deliver higher specificity, while maintaining sufficient sensitivity. Digital mammography, with the opportunity to integrate CAD with image reading, would facilitate such studies. The gain in radiologists' cancer detection rate (when using CAD) compared to the contemporary (probable) increase in recalled, referred and operated healthy women, has to be estimated in order to evaluate CAD as a double-reader.

5.3 INCIDENCE CANCERS (II, IV)

The proportion of reviewed incidence cancers that possibly could be considered to be missed cases varied widely, between 9% and 53%. This variation was due both to the review method used (mixed or non-mixed) and to differences in the number of reviewers/reviewing units (one of them, a majority or all) deemed necessary to identify a case as false negative. When regarding tumours correctly selected by a majority of reviewing units by either method as false negatives, the proportion was found to be 24% (28/117) of the included women, 23% of the 122 tumours. An additional 29% (34/117) of the incidence cancer patients were correctly selected by a minority and may be regarded as "minimal signs". Using this classification, detection rate in this review study is in accordance with the presence of false negatives and/or minimal signs in other studies [79][98][102], though comparisons are difficult to make as there is no strict definition of radiological "minimal signs". As screen-detected cancers often are preceded by radiological "minimal signs" [77][55][79], more often than interval cancers [55], the number of cases accessible on review of prior mammograms can be expected to be high. On the other hand studies show no additional breast cancer risk for

women with mammographically minimal signs when compared to the general screened population [58,59].

If reviewers select women with such "signs" for further assessment recall rate will be high. The reviewers participating in the study obviously differed in their way of dealing with these unspecific radiological signs, demonstrated by the substantial differences in sensitivity and specificity when they were reviewing the mixed set. The unit with highest sensitivity, A, selected a high proportion of observed abnormalities resulting in a high detection rate, more than twice as many as unit B. Unit A also had a high rate of incorrectly selected women (wrong area) and low specificity for healthy women (242 healthy women selected to be recalled by unit A compared to 18 by unit B), as 1 % less specificity meant almost 8 more healthy women recalled. Unit C managed to identify almost as many incidence cancers as unit A but with a higher specificity while unit D had an intermediate result. On the other hand all units showed very high sensitivity to signs typical of malignancy, demonstrated by their ability to detect the included assessable screen-cancers.

To sum up: Reviewers/units do not seem to differ in their ability to detect radiological signs typical of malignancy but in their way of dealing with lesions with a radiological appearance suggesting low probability of cancer.

Despite the high selection rate of some of the units, few additional incidence cancer patients with stage \geq II tumours at presentation, i.e. women who probably would have benefited the most from an earlier detection of their cancer, were correctly selected. Trying to prevent advanced tumours by lowering the threshold for recall was in this retrospective study set-up likely to unnecessarily worry too many healthy women without providing a corresponding gain in sensitivity. The gain in sensitivity due to double reading was of approximately of the same degree as in other studies [72-74].

We found no significant differences regarding tumour size, nodal status, tumour stage, histological subtype nor tumour grade between tumours correctly selected by the reviewers and those not correctly selected. The results did thus not demonstrate women with retrospectively found abnormalities, i.e. with a potentially delayed diagnosis, to more often have an advanced breast cancer diagnosed in the next (presenting) screen than women where no such findings have been made. Accordingly, we noted no tendency for women correctly selected to be overrepresented among women with distant recurrence at follow-up. Even if the result has to be interpreted with caution due to the small number of patients who experienced a distant recurrence, the finding is in agreement with studies correlating prognostic factors to radiological review results [79][98][106] and with previous interval cancer studies [88][93][117].

Just as for interval cancers (I, V) no significant differences in parenchymal density, overall or in the tumour area, nor in tumour location could be demonstrated between malignancies possible to detect compared to those not detectable on review.

Not surprisingly and consistent both with the natural history of the disease and with other reports [50-53][60,61] the number of spiculated lesions and rounded masses increased in the time interval between the preceding and the presenting screen, while

the number of unspecific densities and distortions decreased. Incidence cancers identified in retrospect were more likely to present as spiculated masses and less likely to present as circumscribed lesions and calcifications when compared to incidence cancers not identified, the differences reaching significance. However, there is a substantial overlap with all categories of radiological appearance represented among all the review categories. The overall impression is that no specific radiological feature could be targeted as being representative of tumours possible to detect in retrospect (fig 13) or at stage \geq II at presentation (paper IV), which is in agreement with previous studies [85][118].

Recall rate at the preceding screen was higher for the incidence cancer patients than for the screened population in general, which corresponds with the findings of McCann et al. [118] (7% in the McCann study and 5% in this study). Not surprisingly most false assessments were carried out within the group of false negative tumours as reported by others [128].

Summarizing cases where either a lesion was clearly visible in mammograms prior to the immediately preceding screen, or that had been recalled and (falsely) assessed benign, we regard at least half of the false negative tumours (14 out of 28) as misinterpreted rather than overlooked. This proportion corresponds with previous findings [70] and illustrates that tumours with mammographically slow progression can escape detection due to absence of interval change and thus be misinterpreted as a benign findings or normal breast pattern.

5.4 GENERAL CONSIDERATIONS

Direct comparison of prognosis/patient outcome between incidence and interval cancer patients will be influenced by lead time bias and length bias (se 2:1). As expected and reported in other studies, prognostic factors showed incidence cancers to be less advanced tumours at detection than interval cancers [93][93].

The number of incidence and interval cancers considered to be false negative will be strongly influenced by the classification method used, as demonstrated in the present studies (see also 2:3). Whether healthy women are included in the review set or not and differences in mixing proportions, differences in the number of reviewers required in order to classify a case as missed and whether the reviewers are blinded to patient outcome or not will influence the rate of false negatives.

However differences in the number of retrospectively identified lesions will also be influenced by the study situation, in itself differing from ordinary screening conditions. Some of the reviewers were influenced to select more numerous minor abnormalities than they ordinarily would while others were less affected. Radiologists' unwillingness to worry healthy women with unnecessary recalls, some of whom run the risk of being further referred and subject to unnecessary surgical interventions, is difficult to imitate in a study, while confirmation of professional skill when correctly detecting a malignancy is not. Obviously some reviewers over-read the images, indicated by their study selection rates, highly exceeding their ordinary recall rates (for all units less than 3 %) and with the independent radiologist classifying a few cases selected by a minority in non-mixed reading as too subtle for classification.

To summarize the above mentioned factors: it is difficult to imitate reality in review studies, even with the fixed purpose of keeping the experimental set-up free from undue influences. Nevertheless a mixed set including a high number of healthy women with reviewers blinded to patient outcome will more closely resemble the screening situation than reviews performed with access to the mammograms performed at presentation.

The reviewing of previous mammograms of patients with later diagnosed breast cancer can focus on different aspects. The review could provide an audit tool for improving image interpretation quality. It could also provide a tool for analysing the minimal signs of cancer, studying tumour growth, estimating screening quality or investigating legal aspects. Screening radiologists are probably most interested in obtaining a tool for continuous education with the aim of improving professional skill. Using a review method which will both provide results that reflect the limitations of everyday screening practice and identify areas where this practice can be improved is desirable. A staged method might fulfil this purpose [104]. Furthermore, the use of a majority consensus procedure has the advantage of being independent of the number of reviewers participating and probably therefore more reproducible [80]. As yet no consensus regarding which review method to be used has been established.

Although reviewing previous mammograms of women later diagnosed with breast cancer will hopefully continue to play an important role as learning and teaching tool for both radiologists and radiographers there are more adequate methods of estimation available for quality control. The concept of a successful mammography screening programme as proposed by Tabár et al. [22] in terms of targets to be achieved in order to reduce mortality due to breast cancer, provides more adequate and quicker estimates for quality control than does the rate of retrospectively found false negative cancers.

The substantial differences in detection rate coupled to the image reading circumstances, demonstrated in the present and other studies, is worth considering when the legal aspects of retrospectively found abnormalities attract attention. It is also worth considering that the reduced breast cancer mortality attained in the randomized trials was achieved with the sensitivity of the mammography technology which was in use some 20 years ago. We know that sensitivity of the screen-film system has improved since that time. We also know of reported ways of increasing sensitivity while essentially preserving specificity; double reading with consensus recall decisions [107][72-73], two-view instead of one-view examinations [74-75] and re-screening intervals not exceeding 2 years, preferable shorter for women 40-55 years old [22][76]. Radiologists' conscientious efforts to achieve possible improvements in cancer detection are necessary; still, every additional step to increase sensitivity must be carefully evaluated and adjusted with respect to its impact on specificity. The staffs at the screening units have to communicate this complex insight and the limitations in image cancer detection to the women screened, in order to encourage negative screened women to keep a good lookout for clinical symptoms of breast disease and to return if they develop. Achieving this goal, while preserving the women's confidence in the programme, is a great challenge for the future.

6 CONCLUSIONS

- The review method used and the number of reviewers deemed necessary (one, a majority or all) to classify a case as detectable had a substantial impact on the proportion of interval/incidence cancers that were found to be false negative. When using a method resembling the screen reading situation significantly fewer malignancies were correctly selected by the reviewers than when using a method where radiologists were aware that all images represented cancer patients. Whether the reviewer represented the screening unit responsible for the mammograms or not made no difference to the results.
- Trying to prevent false negative cancers, including advanced tumours, by lowering the threshold for recall was in this retrospective study set-up found likely to unnecessarily worry too many healthy women without providing a corresponding gain in sensitivity.
- Despite sufficiently high sensitivity of CAD alone no increase in radiologist sensitivity, or decrease in specificity, occurred with CAD. Improving CAD specificity, while preserving sensitivity, making radiologists more inclined to revise their interpretations according to CAD, might realise a potential sensitivity increase noted for CAD as a double-reader. Large scale prospective studies would be required to demonstrate the usefulness of radiologists using CAD in every day mass mammography screening.
- Incidence/interval cancer patients with retrospectively found abnormalities, i.e. with a potentially delayed diagnosis, could not be demonstrated to more often have an advanced breast cancer at presentation, or to have a worse outcome at follow-up, than could patients in whom no such findings had been made. No specific radiological feature was found to be representative of false negative malignancies.

7 ACKNOWLEDGEMENTS

I wish to give my warmest thanks to all who have contributed to this work. In particular I would like to express my sincere gratitude to:

Catharina Muren, my supervisor, for your sincere support of my research and for your good advice.

Helene Grundström, my collaborator and co-author, for combining kind support with clear and logical reasoning.

Siv Nilsson, Kerstin Bjurling and all other collaborators at the Department of Mammography, Södersjukhuset, for your support, clear minds and friendship.

Mika Wiege, Hans Lundquist, Eric Havervall, Sven Törnberg, Nils Bjurstam, Brigitte Wilczek, Elisabeth Egge, Lars Rostgaard, Gunilla Svane, Per Sunden, Karin Leifland, Tor Sahlstedt, co-authors, colleagues and collaborators, without your help this study would never been realized.

Margareta Blåsjö and Maria Sylvan, for your invaluable effort in reviewing all the histopathological slides.

Tommy Fornander and Sara Sjöberg-Margolin, my clinical colleagues and collaborators for valuable discussions and together with Toom Singnomklao at the oncologic center, for helping me with data.

Margaretha Höglund, Britta Ågren, Ulla Eriksson and Christina Stjernholm, present and former colleagues at Södersjukhuset, always encouraging my studies.

Göran Elinder, Sari Ponzer, Robert Hahn, Monica Dahlberg, Matts Jonsson, Mia Pettersson and Berit Zanton, the Institution of Södersjukhuset, Karolinska Institutet, for your inexhaustible effort to improve research conditions at Södersjukhuset.

Malin Eriksson for your professional linguistic revision.
Elisabeth Berg for invaluable statistical assistance.

Anders Tyden and Olle Stigwall, Lena Cavallin and Elisabeth Borling, Christina Kilander and Sören Orehag - previously and at present in charge of the Radiology Department Södersjukhuset for encouraging me starting and completing this work.

Medical Library of South Hospital for excellent help and professional advice.

All other collaborators and friends at the Department of Radiology, Oncology, Pathology and Surgery, Södersjukhuset for kindly supporting my studies.

Hans, my husband and friend, Ylva and Björn, my children, for offering me necessary computer assistance and encouraging me even when times were tough.

8 REFERENCES

1. Nyström L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Rydén, 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
2. National Board of Health and Welfare. Cancer incidence in Sweden 2000. Official Statistics of Sweden, Stockholm 2002
3. National Board of Health and Welfare. Causes of Death 2000. Official Statistics of Sweden, Stockholm 2002
4. Carter C, Allen C, Henson D. Relation of tumour size, lymph node status and survival in 24 740 breast cancer cases. *Cancer* 1989; 63: 181-187
5. UICC. TNM classification of malignant tumours, 4th edition. International Union against Cancer. Geneva, 1987
6. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 1991; 19: 403-410
7. Galea M.H. et al. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992; 22(3): 207-219
8. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: An overview of the randomised trials. *Lancet* 1998; 35(9114): 1451-1467
9. Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: An overview of the randomised trials. *Lancet* 1998; 352(9132): 930-942
10. Lundgren B, Jacobsson S. Single view mammography: A simple and effective approach to breast cancer screening. *Cancer* 1976; 38: 1124-1129
11. Lundgren B. The oblique view at mammography. *Br J Radiol* 1977; 50: 626-628
12. Shapiro S, Venet W, Strax Ph. Ten to fourteen year effect of screening on breast cancer mortality. *JNCI* 1982; 60(2): 349-355
13. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977; 39: 2772-2782
14. Shapiro S, Venet W, Strax P et al. Current results from breast cancer screening randomized trial: The Health Insurance Plan (HIP) of Greater New York. In: *Screening for breast cancer*. (Day N, Miller A, eds) p 3-15. H Huber Publishers Toronto 1988
15. Andersson I. Mammographic screening for breast carcinoma, A cross-sectional randomized study of 45-69 year old women. Thesis, 1980 Malmö
16. Andersson I. Breast cancer screening in Malmö. Recent results. *Cancer Res* 1984; 90:114-116
17. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, Ljungberg O, Ranstam J, Sigfússon B. Mammographic screening and

- mortality from breast cancer : the Malmö mammographic screening trial. *B M J* 1988; 297:943-948
18. Tabár L, Gad A. Screening for breast cancer: The Swedish trial. *Radiology* 1981; 138: 219-222
 19. Tabár L, Åkerlund E, Gad A. Five-year experience with single-view mammography randomized controlled screening in Sweden. Recent results. *Cancer Res* 1984; 90:105-113
 20. Tabár L, Fagerberg G, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet* 1985; i: 829-832
 21. Tabár L, Dean P. The control of breast cancer through mammography. What is the evidence? *Radiol Clin North Am* 1987; 25: 993-1005
 22. Tabar L, Fagerberg G, Duffy S, Day N, Gad A, Gröntoft O. Update of Swedish two-county program of mammographic screening for breast cancer. *Radiol. Clin. North Am.* 1992; 30: 187-210
 23. Bjurstam N, Björnelid L, Duffy SW, et al. The Gothenburg breast screening trial. First results on mortality, incidence and mode of detection for women ages 39-49 years at randomization. *Cancer* 1997; 80: 2091-2099
 24. Roberts M, Alexander F, Andersson T et al. Edinburgh trial of screening for breast cancer: mortality at 7 years. *Lancet* 1990; 335: 241-246
 25. Frisell J, Eklund G, Hellström L, Lidbrink E, Rutqvist L-E, Somell. A Randomized study of mammography screening - preliminary report on mortality in the Stockholm trial *Breast Cancer Res Treat* 1991; 18:49-56
 26. Frisell J, Lidbrink E, Hellström L, Rutqvist L-E. Followup after 11 years - update of mortality results in the Stockholm mammographic screening trial *Breast Cancer Res Treat* 1997; 45: 263-270
 27. Official recommendation from The National Board of Health and Welfare: Mammographic screening for early detection of breast cancer. *Socialstyrelsen, Stockholm* 1986 and 1990.
 28. Tabár L, Vitak B, Chen Tony H, et al. Beyond randomized controlled trials. Organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; 91:1724-1731
 29. Thurfjell E, Lindgren A Population-based mammography screening in Swedish clinical practice: prevalence and incidence screening in Uppsala county *Radiology* 1994; 193:351-357
 30. Lidbrink E, Frisell J, Brandberg Y, et al. Non-attendance in the Stockholm mammography screening trial: relative mortality and reasons for non-attendance. *Breast Cancer Res Treat* 1999; 35: 267-275
 31. UK Trial of Early Detection of Breast Cancer Group. First results on mortality reduction in the UK trial of early detection of breast cancer. *Lancet* 1988; ii: 411-416
 32. European Commission. European Guidelines for quality assurance in mammography screening. *Public Health 2nd, ECSC-EC-EAEC Brussels* 1996
 33. Lidbrink E, Törnberg S, Azavedo E, Frisell J, Hjalmar M-L, Leifland K, Sahlstedt T, Skoog L. The general mammography screening programme in Stockholm *Acta Onc* 1994; 33(4): 353-358

34. Miller A, Baines C, To T, et al. Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Can Med Assoc J* 1992; 147: 1459-1468.
35. Miller A, Baines C, To T, et al. Canadian national breast screening study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Can Med Assoc J* 1992; 147: 1477-1488.
36. Lidbrink E, Levi L, Pettersson I, Rosendahl I, Rutqvist LE, de la Torre B, Wasserman J, Wiege M. Single-view screening mammography: psychological, endocrine and immunological effects of recalling for a complete three-view examination. *Eur J Cancer* 31 A, No 6: 932-933, 1995
37. Lidbrink E, Elfving J, Frisell J, Jonsson E. Neglected aspects of false positive findings of mammography in breast cancer screening: analysis of false positive cases from the Stockholm trial. *B M J* 312: 273-276, 1995
38. Gram I, Lund E, Slenker S. Quality of life following a false positive mammogram. *Br J Cancer* 1990; 62: 1018-1022
39. Dixon J, John T. Morbidity after breast biopsy for benign disease in a screened population. *Lancet* 1992; 339: 128.
40. Wolfe JN. Breast parenchymal patterns and their changes with age. *Radiology* 1976; 121(3 Pt. 1):542-52
41. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. *AJR* 1976; 126 1130-1139
42. Tabár L, Dean PB. Teaching atlas of mammography, 2nd ed. New York: Thieme Inc, 1985
43. Tabár L Viewing technique. Interdisciplinary breast conference. Radiological Society of Finland Helsinki Nov 9-11 2000
44. Hanley R. Carcinoma of the breast. *Ann R Coll Surg Engl* 1975; 57: 59-66
45. Brown M, Eccles C, Wallis M. Geographical distribution of breast cancers on the mammogram: an interval cancer database. *Br J Radiol* 2001; 74: 317-322
46. Georgen S, Evans J, Cohen G et al. Characteristics of breast carcinomas missed by screening radiologist. *Radiology* 1997; 204: 131-135
47. Hendrick E. Mammography quality assurance. *Current Issues. Cancer* 1993; 72: 1466-1474
48. Hoffman F, Rheinstein P, Houn F. The mammography quality standard act of 1992. *Am Fam Phys* 1994; 49: 1965-1970
49. Sickles E. Quality assurance. How to audit your own mammography practice. *Radiol Clin North Am* 1992; 30: 265-275
50. Warren R, Young J, McLean L et al. Radiology review of the UKCCR breast screening frequency trial: Potential improvements in sensitivity and lead time of radiological signs. *Clin Radiol* 2003; 58: 128-132
51. Meeson S, Young K, Wallis M et al. Image features of true positive and false negative cancers in screening mammograms. *Br J Radiol* 2003; 76: 13-21
52. Ikeda D, Birdwell R, O'Shaughnessy K, et al. Analysis of 172 subtle findings on prior normal mammograms in women with breast cancer detected at follow-up screening. *Radiology* 2003; 226: 494-503

53. Burell H, Pinder S, Wilson R, et al. The positive predictive value of mammographic signs: A review of 426 non-palpable breast lesions. *Clin Radiol* 1996; 51: 277-281
54. Hilleren D, Andersson I, Lindholm K, et al. Invasive lobular carcinoma: mammographic findings in a 10-year experience. *Radiology* 1991; 178: 149-154
55. van Dijck J, Verbeek A, Hendriks J, et al. The current detectability of breast cancer in a mammographic screening program. *Cancer* 1993; 72:1933-1938
56. Harvey J, Fajardo L, Innis C. Previous mammograms in patients with impalpable breast carcinoma: Retrospective vs blinded interpretation. *AJR* 1993; 161:1167-1172
57. Duncan K, Needham G, Gilbert F, et al. Incident round cancers: What lessons can we learn? *Clin Radiol* 1998;53:29-32
58. Maes R, Dronkers D, Hendriks J, Thijssen M, Nab H. Do non-specific minimal signs in a biennial mammographic breast cancer screening programme need further diagnostic assessment? *BJR* 1997;70:34-38
59. Wolverson D, Sickles E. Clinical outcome of doubtful mammographic findings. *AJR* 1996; 167: 1041-1045
60. Kopans D. *Breast Imaging*. 2nd ed. 1998 Lippincott-Raven Publishers
61. Svane G, Potchen J, Sierra A & Azavedo E. *Screening Mammography, Breast Cancer Diagnosis in Asymptomatic Women*. E. Editors, 1993 Mosby
62. Feig S, Shaber G, Patchefsky A, et al. Tubular carcinoma of the breast *Radiology* 1978; 129: 311
63. Ciatto S, Roselli del Turco M, Bonardi R, et al. An analysis of the characteristics of the interval cancers in the Florence district programme 1989-91. *Radiol Med (Torino)* 1994; 87 412-416
64. Azavedo E, Svane G. Radiological aspects of breast cancers detected through a breast cancer screening program. *Eur J Radiol* 1991; 13: 88-92
65. Tabár L, Chen H-H, Duffy SW, et al. A novel method for prediction of long-term outcome of women with T1a, T1b and 10-14 mm invasive cancers: a prospective study. *Lancet* 2000; 355: 429-433
66. Thurfjell E, Thurfjell MG, Lindgren A. Mammographic finding as a predictor of survival in 1-9 mm invasive breast cancers. Worse prognosis for cases presenting as calcifications alone. *Breast Cancer Res Treat* 2001; 67: 177-180
67. Baker J, Kornguth P, Floyd C. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR* 1996; 166: 773-778
68. Vyborny C, Doi T, O'Shaughnessy K et al. Breast cancer: Importance of spiculation in computer-aided detection. *Radiology* 2000; 215: 703-707
69. Tabár L, Dean P, Kaufman C, Duffy S, Chen H-H. A new era in the diagnosis of breast cancer. *Surg Onc Clin North Am* 2000; 9: 233-277
70. Bird R, Wallace T, Yankaskas B. Analysis of cancers missed at screening mammography. *Radiology* 1992; 184:613-617

71. Meyer J, Kopans D. Stability of a mammographic mass: a false sense of security. *AJR* 1981; 137: 595-598
72. Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films - one radiologist or two? *Clin Radiol* 1993; 48: 414-421
73. Thurfjell E, Lernevall K, Taube A. Benefit of independent double reading in a population- based mammography screening program. *Radiology* 1994; 191:241-4
74. Wald N, Murphy P, Major P, et al. UKCCCR multicentre randomized controlled trial of one and two view mammography in breast cancer screening. *BMJ* 1995; 311: 1189-1193
75. Warren R, Duffy S, Bashir S, et al. The value of the second view in screening mammography. *Br J Radiol* 1996 ;69; 105-108
76. Tabár L, Fagerberg G, Day N, et al. 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
77. Harvey J, Fajardo L, Innis C. Previous mammograms in patients with impalpable breast carcinoma: Retrospective vs blinded interpretation. *AJR* 1993; 161:1167-1172
78. Duncan A, Wallis M. Classifying Interval Cancers. *Clin Radiol* 1995;50:774-7
79. Duncan K, Needham G, Gilbert F, et al. Incident round cancers: What lessons can we learn? *Clin Radiol* 1998;53:29-32
80. Britton P, McCann J, O'driscoll D, et al. Interval cancer peer review in East Anglia: Implications for monitoring doctors as well as the NHS breast screening programme. *Clin Radiol* 2000; 56: 44-49
81. Gower-Thomas K, Fielder H, Branston L, et al. Reviewing interval cancers: Time well spent? *Clin Radiol* 2002; 57: 384-388
82. Lundgren B. Efficiency of single-view mammography. Rate of interval cancer cases. *J Natl Cancer Inst* 1979;62:799-803
83. Holmberg L, Tabar L, Adami H, Bergström R. Survival in breast cancer diagnosed between mammographic screening examinations. *Lancet* 1986;II:27-30
84. Vitak B, Stål O, Månsson B A, et al. Interval cancers and cancers in non-attenders in the Östergötland Mammographic screening Programme. Duration between screening and diagnosis, S-face fraction and distant recurrence. *Eur J Cancer* 1997; 33(9):1453-60.
85. Vitak B, Invasive interval cancers in the Östergötland mammographic screening programme. Radiological analysis. *European Radiology* 1998
86. Ikeda D, Andersson I, Wattsgård C, et al. Interval carcinomas in the Malmö mammographic screening trial. *AJR* 1992; 159:287-294
87. Andersson I. What can we learn from interval carcinomas. *Recent Results in Cancer Res* 1984; 90: 161-164

88. Frisell J, von Rosen A, Wiege M, et al. Interval cancers and survival in a randomized breast cancer screening trial in Stockholm. *Breast Cancer Res Treat* 1992; 24:11-16.
89. Otten J, Dijck van J, Peer P, et al. Long term breast cancer screening in Nijmegen, The Netherlands: the nine rounds from 1975-92. *J Epidemiol Community Health* 1996; 50:353-358
90. Burhenne J, Warren Burhenne L, Goldberg F, et al. Interval breast cancers in the screening mammography program of British Columbia: Analysis and classification. *AJR* 1994; 162: 1067-1071
91. Holland R, Mravunac M, Hendriks J, Bekker B. So-called interval cancers of the breast. *Cancer* 1982; 49: 2527-2533
92. Peeters P, Verbeek A, Hendriks J, Holland R, Mravunac M. The occurrence of interval cancers in the Nijmegen screening program. *Br J Cancer* 1989;59:929-932
93. Brekelmans C, Peeters P, Deurenberg J, Collette H. Survival in interval breast cancer in the DOM screening programme. *Europ J Cancer* 1995;31A 11:1830-1835
94. Martin J, Moskowitz M, Milbrath J. Breast cancer missed by mammography. *AJR* 1979; 132: 737-739
95. Burell H, Sibbering M, Wilson R, et al. Screening interval breast cancers: Mammographic features and prognostic factors. *Radiology* 1996; 199:811-817
96. Everington D, Gilbert F, Tyack C, et al. The Scottish breast screening programme's experience of monitoring interval cancers. *J Med Screen* 1999; 6: 21-27
97. Sylvester P, Vipond M, Kutt E, et al. Rate and classification of interval cancers in the breast screening programme. *Ann R Coll Surg Engl* 1997; 79: 276-277
98. Maxwell A. Breast cancers missed in the prevalent screening round: effect upon the size distribution of incident round detected cancers. *J Med Screen* 1999; 6: 28-29
99. Saarenmaa I, Salminen T, Geiger U, et al. The visibility of cancer on earlier mammograms in a population-based screening programme. *Eur J Cancer* 1999; 35: 1118-1122
100. Daly C, Apthorp L, Field S. Second round cancers: How many were visible on the first round of the UK National Breast Screening Programme, three years earlier? *Clin Radiol*; 53:25-28
101. Day N, McCann J, Camilleri-Ferante C, Britton P, Hurst G, Cush S, Duffy S. Monitoring interval cancers in breast screening programmes: the east Anglian experience. Quality Assurance Management Group of the East Anglian Breast Screening Programme. *J Med Screen* 1995; 2(4): 180-185
102. Jones R, McLean L, Young J, et al. Proportion of cancers detected at first incident screen which were false negative at the prevalent screen. *The Breast* 1996; 5:339-343
103. Simpson W, Neilson F, Young J R. The identification of false negatives in a population of interval cancers: a method for audit of screening mammography. *The Breast* 1995;4:183-8

104. Amos A, Kavanagh A, Cawson J. Radiological review of interval cancers in an Australian mammographic screening programme. *J Med Screen* 2000; 7: 184-189
105. Rijke J, Schouten L, Schreutelkamp J, et al. A blind review and an informed review of interval breast cancer cases in the Limburg screening programme, the Netherlands. *J Med Screen* 2000; 7: 19-23
106. Wheatley D, Yeoman L, Burell H, et al. Mammographic and pathological features of breast cancer detected at first incident round screening. *The Breast* 1997; 6: 259-265
107. Warren RML. & Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *Br J Radiol* 1995; 68; 958-962.
108. Giger M, MacMahon H. Image processing and computer-aided diagnosis. *North Am Clin Radiol* 1996; 3; 565-596.
109. Vyborny CJ, Giger M. Computer vision and artificial intelligence in mammography. *A J R* 1994; 162; 699-708.
110. Kegelmeyer W P, Pruneda J M, Bourland P D, Hillis A, Riggs M W, Nipper M L. Computer-aided mammographic screening for spiculated lesions. *Radiology* 1994; 191:331-337
111. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y, Macmahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms. *Investigative Radiology* 1990; 25:1102-1110
112. Thurffjell E, Thurffjell M.G, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. *Acta Radiologica* 1998; 39:384-388
113. Warren Burhenne L, Wood S, D'Orsi C et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000; 215: 554-562
114. Birdwell R, Ikeda D, O'Shaughnessy K et al. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology* 2001; 219: 192-202
115. Warren Burhenne L, Laurie L, O'Shaughnessy K et al. Invasive lobular carcinoma of the breast: Mammographic characteristics and computer-aided detection. *Radiology* 2002; 225: 182-189
116. Freer T, Ulissey M. Screening mammography with computer-aided detection: Prospective study of 12 860 patients in a community breast center. *Radiology* 2001; 220: 781-786
117. Vitak B. Interval cancers, positive predictive value for malignancy, and other early quality indicators in mammographic screening for breast cancer. Thesis Linköping University and Karolinska Institute, 1998 Linköping
118. McCann J, Britton P, Warren R, et al. Radiological peer review of interval cancers in the East Anglian breast screening programme: what are we missing? *J Med Screen* 2001; 8:77-85

119. Brekelmans C T, van Gorps J M, Peeters P H, Colette H J. Histopathology and growth rate of interval breast carcinoma. Characterization of different subgroups. *Cancer* 1996; Sep 15; 78(6): 1220-1228
120. von Rosen A, Erhardt K, Hellström L, Somell A, Auer G. Assessment of malignancy potential in so-called interval mammary carcinomas. *Br Canc Res Treatm* 1985;6:221-227
121. Schroen A, Wobbes T, van Der Sluis R. Interval carcinomas of the breast: A group with intermediate outcome. *J Surg.Oncol.* 1996;63:141-144
122. Mandelsohn M, Oestreicher N, Porter P et al. Breast density as a predictor of mammographic detection: Comparison of interval- and screen-detected cancers. *J Natl Cancer Inst* 2000; 92, 13: 1081-1087
123. Sala E, Warren R, McCann J et al. Mammographic parenchymal patterns and mode of detection: Implications for the breast screening programme. *J Med Screen* 1998; 5: 207-212
124. Byng J, Yaffe M, Jong R, et al. Analysis of mammographic density and breast cancer risk from digitized mammograms. 1998 *Radiographics*; 18:1587-1598
125. van Gils C, Otten J, Verbeek A, et al. Effect of mammographic breast density on breast cancer screening performance: A study in Nijmegen, the Netherlands. *J Epidemiol Commun Health* 1998; 52:267-271
126. Boyd N, Lockwood G, Martin L et al. Mammographic densities and risk of breast cancer among subjects with a family history of this disease. *J Natl Cancer Inst* 1999; 91: 1404-1408
127. Feig S, Shaber G, Patchefsky A et al. Analysis of clinically occult and mammographically occult breast tumours. *AJR* 1977; 128: 403-408
128. Burell H, Evans A, Wilson R, et al. False-negative breast screening assessment. What lessons can we learn? *Clin Radiol* 2001; 56: 385-388
129. Warren R, Duffy S. Interval cancers as an indicator of performance in breast screening. *Breast Cancer* 2000; 7: 9-18
130. Bröstcancergruppen och Onkologiskt Centrum för Stockholm-Gotlandregionen, Vårdprogram för bröstcancer Onkologiskt Centrum Karolinska Sjukhuset 1986
131. WHO. International histological classification of tumours. Histological typing of breast tumours. WHO. Geneva, 1981
132. Linell F, Ljungberg O, Andersson I Breast Carcinoma Aspects of early stages, progression and related problems. 1980 Munksgaard Copenhagen
133. Altman D. Practical Statistics for Medical Research. 1991 Chapman&Hall.
134. Siegel, S. and Castellan, N. J. Nonparametric Statistics for the Behavioural Sciences. 1989 New York: McGraw-Hill