Randomized Trial of Screen-Film versus Full-Field Digital Mammography with Soft-Copy Reading in Population-based Screening Program: Follow-up and Final Results of Oslo II Study[1]

Per Skaane, MD, PhD
Solveig Hofvind, PhD
Arnulf Skjennald, MD, PhD

Purpose: To prospectively compare performance indicators at screen-film mammography (SFM) and full-field digital mammography (FFDM) in a population-based screening program.

Materials and Methods: The regional ethics committee approved the study; informed consent was obtained from patients. Women aged 45–69 years were assigned to undergo SFM ($n = 16\,985$) or FFDM ($n = 6944$). Two-view mammograms were interpreted by using independent double reading and a five-point rating scale for probability of cancer. Positive scores were discussed at consensus meetings before decision for recall. The group was followed up for 1.5 years (women aged 45–49 years) and 2.0 years (women aged 50–69 years) to include subsequent cancers with positive scores at baseline interpretation and to estimate interval cancer rate. Recall rates, cancer detection, positive predictive values (PPVs), sensitivity, specificity, tumor characteristics, and discordant interpretations of cancers were compared.

Results: Recall rate was 4.2% at FFDM and 2.5% at SFM ($P < .001$). Cancer detection rate was 0.59% at FFDM and 0.38% at SFM ($P = .02$). There was no significant difference in PPVs. Median size of screening-detected invasive cancers was 14 mm at FFDM and 13 mm at SFM. Including cancers dismissed at consensus meetings, overall true-positive rate at baseline reading was 0.63% at FFDM and 0.43% at SFM ($P = .04$). Sensitivity was 77.4% at FFDM and 61.5% at SFM ($P = .07$); specificity was 96.5% and 97.9%, respectively ($P < .005$). Interval cancer rate was 17.4 at FFDM and 23.6 at SFM. The proportion of cancers with discordant double readings was comparable at FFDM and SFM.

Conclusion: FFDM resulted in a significantly higher cancer detection rate than did SFM. The PPVs were comparable for the two imaging modalities.

[1] From the Department of Radiology, Ullevaal University Hospital, University of Oslo, Kirkeveien 166, NO-0407 Oslo, Norway (P.S., A.S.); and Institute of Population-based Cancer Research, The Cancer Registry of Norway, Oslo, Norway (S.H.). From the 2003 RSNA Annual Meeting. Received August 25, 2006; revision requested October 25; revision received November 14; final version accepted January 3, 2007. **Address correspondence to** P.S. (e-mail: *per.skaane@ulleval.no*).

Four large-scale studies (1–4) that involved comparison of screen-film mammography (SFM) and full-field digital mammography (FFDM) by using soft-copy reading in breast cancer screening have been published to date. Results of the Colorado-Massachusetts study (1,5) showed a nonsignificantly higher cancer detection rate at SFM than at FFDM, with a lower recall rate at FFDM as the only statistically significant finding. The second trial, the Oslo I study (2), was a paired study that included 3683 women invited to participate in population-based screening program. Results of this study showed a lower but statistically nonsignificant cancer detection rate at FFDM. Analysis of false-negative FFDM interpretations and final results based on 2 years of follow-up from the Oslo I study were published in 2005 (6). The third published trial, the Oslo II study (3), was a randomized trial in a population-based screening program. In this study, FFDM had a higher cancer depiction rate than did SFM, with a difference that approached statistical significance. The fourth study, the Digital Mammographic Imaging Screening Trial (DMIST), was a paired multicenter study that involved the application of different digital systems in the examination of 49 528 women (4). Results of this study demonstrated that the overall diagnostic performance at SFM and FFDM was comparable, but there was a significantly higher performance at FFDM in women younger than age 50 and in women with radiographically dense breast parenchyma (4).

Mammograms of cancers with discordant findings at double reading constitute an important subset of examination findings in breast screening programs. Tumors manifesting as interval cancers or as cancers at subsequent screening rounds might be dismissed at baseline consensus meetings or panel arbitration (6,7). Cancers might also be missed at diagnostic work-up because of interpretation error or because of sampling errors at needle biopsy and might manifest as subsequent cancers. Failure to act on these nonspecific mammographic findings prospectively does not necessarily constitute interpretation below a reasonable standard of care (8).

The purpose of our study was to prospectively compare performance indicators at SFM and FFDM in a population-based screening program.

Materials and Methods

Study Group

The study was approved by the regional ethical committee, and all women gave informed consent. Women assigned to undergo FFDM were informed about this in their invitation letter, and their participation in the study was voluntary. Women included in this study were invited by a personal letter to participate in the population-based Norwegian Breast Cancer Screening Program (NBCSP) in Oslo.

Baseline mammographic examinations were performed between November 27, 2000, and December 31, 2001. Women aged 50–69 years were part of the NBCSP, which conducts biennial screening, whereas women aged 45–49 years were offered annual screening in Oslo County. The randomization process started on October 26, 2000. Stratified randomization according to age and residence was performed by the Norwegian National Health Screening Service. It was decided that about 70% of invited women should be assigned to undergo SFM because of equipment availability (two SFM laboratories and one FFDM laboratory were available) (3). Invitation letters were mailed to the women about 3–4 weeks in advance. If a woman did not attend the screening examination, a reminder letter was mailed 4–8 weeks after the scheduled time.

Rechecking the screening database revealed that some women attended the program after having received a reminder letter that followed the initial invitation for an examination date before the randomization started. These women had erroneously been included in the analysis in our preliminary report because the examination dates were within the study period; how-

Advances in Knowledge

- Full-field digital mammography (FFDM) enabled a significantly ($P = .03$) higher cancer detection rate than did screen-film mammography (SFM).
- There was no significant difference ($P = .68$) in positive predictive value between FFDM and SFM.
- The interval cancer rate was lower at FFDM than at SFM, but the difference was not significant ($P = .35$).
- The proportion of cancers with discordant double reading was comparable for the two imaging modalities (25% at FFDM and 30% at SFM).

Implications for Patient Care

- The higher cancer detection rate at full-field digital mammography (FFDM) might improve the detection of small cancers in breast cancer screening programs, thus making it possible for more women to be treated with breast-conserving surgery.
- The higher cancer detection rate and the lower interval cancer rate at FFDM might contribute to a reduction in mortality from breast cancer.

Published online
10.1148/radiol.2443061478

Radiology 2007; 244:708–717

Abbreviations:
DMIST = Digital Mammographic Imaging Screening Trial
FFDM = full-field digital mammography
NBCSP = Norwegian Breast Cancer Screening Program
PPV = positive predictive value
SFM = screen-film mammography

Author contributions:
Guarantor of integrity of entire study, P.S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; literature research, P.S., A.S.; clinical studies, P.S., A.S.; statistical analysis, P.S., S.H.; and manuscript editing, all authors

Authors stated no financial relationship to disclose.

ever, they were excluded from analysis in our Oslo II follow-up study. Women who attended after receiving a reminder letter that followed the initial invitation in the study period had been properly assigned and were thus included in the study. Women undergoing scheduled screening examinations before December 31, 2001 but undergoing their diagnostic work-up in the beginning of 2002 were all included in the study group.

Women examined at the Breast Imaging Center, Ullevaal University Hospital, instead of the screening unit were excluded from analysis because independent double reading could not be guaranteed for mammograms in these women. Furthermore, it was decided that the date of examination, and not the date of randomization (as in the preliminary Oslo II study), should be used for categorizing the women into the age groups according to the guidelines of the NBCSP. Thus, the study population consisted of 23 929 women, of whom 13 912 were in the age group 50–69 years and 10 017 were in the age group 45–49 years. A total of 16 985 women underwent SFM, and 6944 women underwent FFDM.

Imaging

SFM examinations were performed with one of two units (Mammomat 300; Siemens Medical Systems, Erlangen, Germany) with Min-R 2000 film and Min-R 2190 screens (Eastman Kodak, Rochester, NY) in both standard and large formats. FFDM images were acquired with another unit (Senographe 2000D; GE Medical Systems, Milwaukee, Wis). Mammograms from both imaging modalities (SFM and FFDM) included the two standard views (craniocaudal and mediolateral oblique) of each breast. Further details of imaging procedures were presented in our preliminary report (3).

Image Interpretation

Images were interpreted the following day in a batch reading. Eight radiologists (including P.S. and A.S.) with 4–10 years of experience in screening mammography participated in image interpretation during the study period. All readers had taken part in the Oslo I study and were consequently experienced in both SFM and FFDM with soft-copy reading. Independent double reading was the standard for both SFM and FFDM. Prior mammograms were not used in the interpretation session but were always offered at both the SFM and FFDM consensus meetings, if available. SFM images were read by using two standard motorized mammography alternators, and FFDM images were read by using soft-copy reading at a review workstation (GE Medical Systems) that included two high-resolution 2000 × 2500-pixel monitors and a dedicated keypad. Results of SFM and FFDM readings were recorded directly into the database of the Norwegian Cancer Registry by using a light pen (bar code technology) or a mouse.

A five-point rating scale for the probability of cancer was used for interpretation of both SFM and FFDM images, as follows: a score of 1 indicated normal or definitely benign findings; a score of 2 indicated probably benign findings; a score of 3 indicated indeterminate finding; a score of 4 indicated probably malignant findings; and a score of 5 indicated malignant findings. If at least one of the two readers categorized a mammographic finding with a score of 2 or higher (defined as a positive score), the mammograms were automatically selected for discussion at the consensus meeting. Details of image interpretation and hanging protocol were described in our preliminary report (3).

Consensus (arbitration) meetings were held twice a week, and all radiologists were encouraged to participate. Usually, only the prescribed minimum of two radiologists were present. The radiologists who attended the consensus meetings were not necessarily the same readers who interpreted the SFM or FFDM images. The outcome of the consensus meeting was a decision about which women should be dismissed and go back to the screening program and which should be recalled for diagnostic work-up. At the consensus meeting, radiologists were free to dismiss cases with abnormal findings categorized with scores no higher than 2 by one or both readers. Short-term follow-up is not used in our screening program (3).

Diagnostic Work-up

Diagnostic work-up was performed in a single visit within 2 weeks after the consensus meeting. Work-up included the acquisition of spot-compression and magnification views, ultrasonographic (US) images, and magnetic resonance images, if needed. Fine-needle aspiration cytology was the technique used for percutaneous biopsies. Imaging-guided biopsies were performed by using US guidance whenever possible; otherwise, stereotactic guidance was used. Cytologic and histologic examinations were performed in the Department of Pathology, Ullevaal University Hospital, and all results were reported to the Norwegian Cancer Registry.

Follow-up and Subsequent Cancers

The study group aged 50–69 years was followed up for 2 years to include interval cancers and cancers discovered at the subsequent screening round with a true-positive baseline interpretation in the comparison of the two modalities. An unexpected challenge occurred in 2002 when the Norwegian government decided to stop the screening program for women aged 45–49 years, and follow-up is thus limited to 18 months for this age group. Cancers diagnosed after a negative (normal) finding at baseline screening examination (score of 1 by both independent readers) or cancers that had a true-positive score but were dismissed at consensus meeting or diagnostic work-up and were diagnosed before the next scheduled screening examination were all defined as interval cancers. Results of baseline screening examinations, results of diagnostic work-up in recalled women, and subsequent cancers are registered in a nationwide screening database in the Norwegian Cancer Registry. Almost 100% surveillance of the study population is possible through record linkage of the screening database of the NBCSP to the national cancer database by use of an 11-

digit identification number given to all Norwegian inhabitants. Since 1953, it has been mandatory by law to report all cancer cases to the Norwegian Cancer Registry.

Statistical Analysis

Medical audit parameters (performance indicators) for mammographic screening, including recall rate, cancer detection rate, and positive predictive value (PPV), were calculated for both modalities. PPV_1 was the proportion of cancers among the women recalled, while PPV_2 was the proportion of cancers among women who underwent needle biopsy at initial work-up. The interval cancer rate was defined as the number of interval cancers diagnosed since the last negative finding at a screening examination per 10 000 women with a negative finding at screening. The χ^2 test was used to compare recall rates, PPVs, and cancer detection rates. *P* values less than .05 were considered to indicate a statistically significant difference. The observer agreement for cancers for two independent readers was evaluated by using κ statistics. Analyses were conducted by using software (Epi Info, version 6, Centers for Disease Control and Prevention, Atlanta, Ga; SPSS, version 12.0.1 for Windows, SPSS, Chicago, Ill).

Figure 1

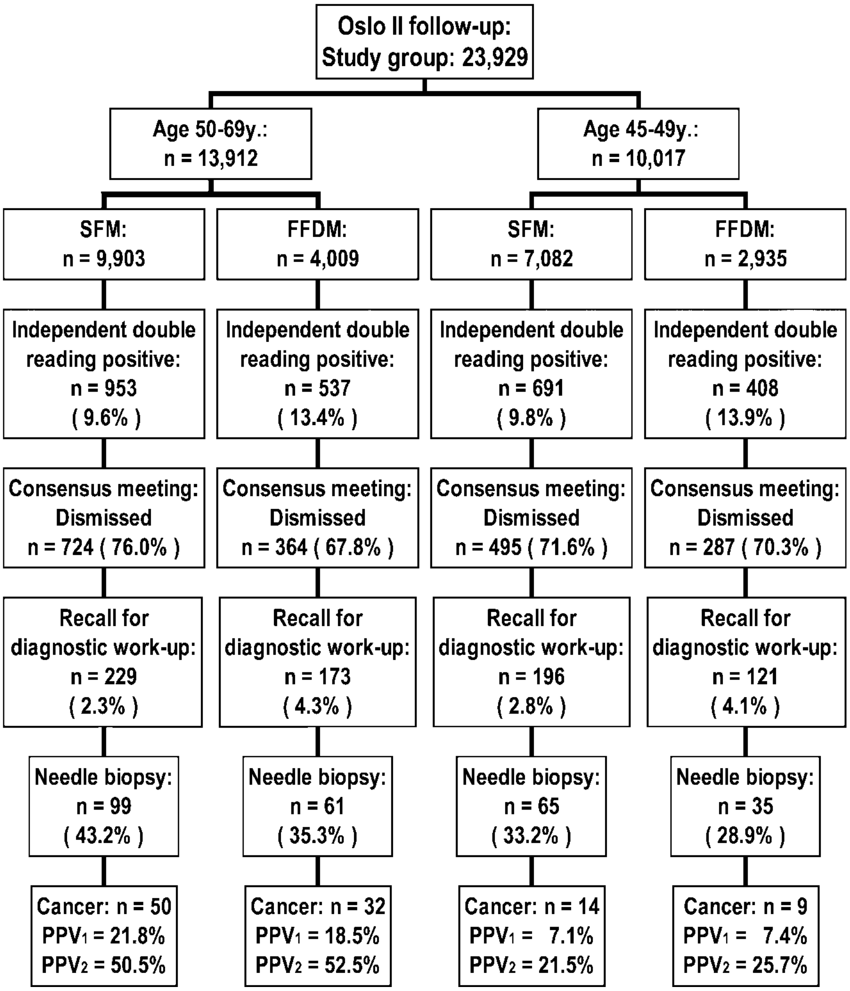

Figure 1: Flowchart shows numbers of the following: women undergoing SFM and FFDM, positive scores at independent double reading, dismissed cases, women recalled for diagnostic work-up, women who underwent needle biopsy at diagnostic work-up, and screening-detected cancers according to age group. PPV_1 = proportion of cancers among women recalled for diagnostic work-up. PPV_2 = proportion of cancers among women who underwent needle biopsy.

Results

Positive Scores and Recall Rates

Positive interpretation at independent double reading (score of 2 or higher assigned by at least one of the two readers) occurred in 953 (9.6%) of 9903 cases at SFM and in 537 (13.4%) of 4009 cases at FFDM in age group 50–69 years (Fig 1). For age group 45–49 years, positive interpretation occurred in 691 (9.8%) of 7082 cases at SFM and in 408 (13.9%) of 2935 cases at FFDM. Thus, the overall (both age groups) positive reading at baseline interpretation was 945 (13.6%) of 6944 cases at FFDM and 1644 (9.7%) of 16 985 cases at SFM ($P < .01$).

A total of 1219 (74.1%) of 1644 cases with a positive score at baseline interpretation at SFM were dismissed at consensus meeting; 651 (68.9%) of 945 such cases at FFDM were dismissed ($P = .004$). The recall rate for both age groups was 294 (4.2%) of 6944 cases at FFDM and 425 (2.5%) of 16 985 cases at SFM ($P < .01$).

Detection Rate of Baseline Cancers

A total of 120 screening-detected cancers were diagnosed during the study period, of which six were excluded from analysis in our preliminary report (3). Our scrutiny of the screening database for subsequent cancers revealed that one screening-detected cancer had been overlooked in our preliminary report. This woman in age group 45–49 years who underwent FFDM examination had a delay of more than 7 months from screening examination to cancer diagnosis because of sampling error at the first needle biopsy and unsuccessful first preoperative localization. The woman was, however, never dismissed from work-up; consequently, this cancer was by definition a screening-detected cancer. Ten cancers diagnosed among women who attended the screening examination after having received a reminder letter and women who were examined at the Breast Imaging Center were excluded from further analysis; the 10 cancers included the following: six cancers at SFM in age group 50–69 years (five women who received a reminder letter and one disabled woman who was examined at the Breast Imaging Center), three cancers at SFM in age group 45–49 years (all women who received a reminder letter), and one cancer at FFDM in age group 50–69 years (woman with prosthesis examined at the Breast Imaging Center). Thus, a total of 105 baseline screening-detected cancers were included for further analysis (Fig 1).

The overall (both age groups) cancer detection rate was 41 (0.59%) of 6944 cases at FFDM and 64 (0.38%) of 16 985 cases at SFM (*P* = .03) (Fig 2).

Figure 2

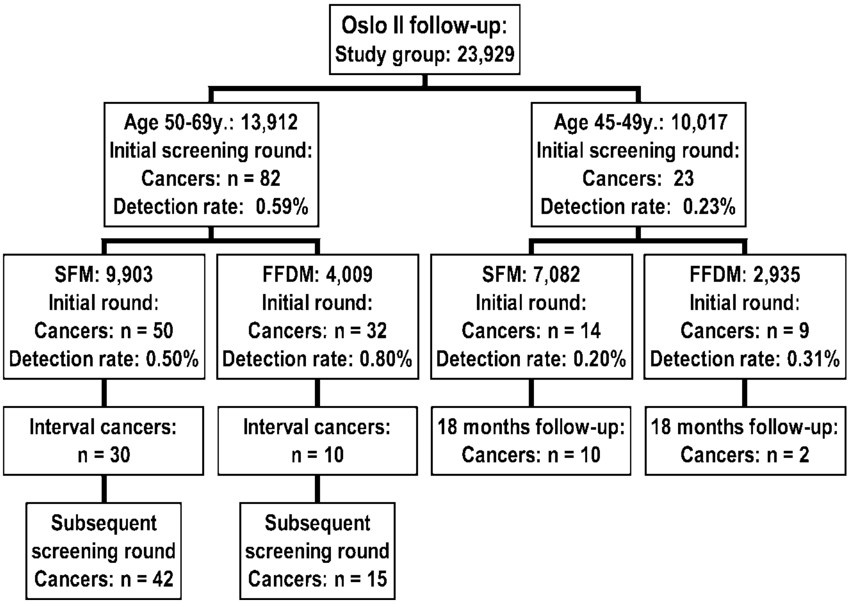

Figure 2: Flowchart shows numbers of cancers detected at baseline screening, interval cancers, and subsequent cancers according to age group. For age group 45–49 years, follow-up was 18 months because the program was stopped by Norwegian government during the study period.

Table 1

Histologic Findings and Size of Cancers Diagnosed at SFM and FFDM at Baseline Screening

Histologic Finding	Cancers Detected at SFM	Cancers Detected at FFDM
DCIS	20 (31)	11 (27)
Invasive carcinomas	44 (69)	30 (73)
IDC with or without DCIS*	36	26
ILC with or without DCIS	7	3
Other carcinomas	1	1
Size		
Not measurable*	1	0
Mean size (mm)	14.0	14.9
Median size (mm)	13.0	14.0
Total	64	41

Note.—Unless otherwise indicated, data are numbers of cancers, with percentages in parentheses. DCIS = ductal carcinoma in situ, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma.

* One SFM-detected cancer finding confirmed at needle biopsy (invasive ductal carcinoma) showed total regression after neoadjuvant chemotherapy. This cancer is excluded from mean and median tumor size.

The percentages of screening-detected ductal carcinoma in situ cases and invasive carcinomas and mean and median size of invasive carcinomas detected with the two imaging modalities were comparable (Table 1). Comparison of cancer detection rates according to 5-year intervals of the age range of the two groups showed a higher detection rate at FFDM in nearly all groups, and FFDM had a significantly higher cancer detection rate for all cancers and for invasive cancers, but not for ductal carcinoma in situ (Table 2).

PPVs

Overall (both age groups), PPV_1 was 64 (15.1%) of 425 cases at SFM and 41 (13.9%) of 294 cases at FFDM (*P* = .68) (Fig 1). The corresponding overall PPV_2 was 64 (39.0%) of 164 cases at SFM and 41 (42.7%) of 96 cases at FFDM (*P* = .56).

Interval and Subsequent Screening Round Cancers

A total of 109 subsequent cancers were diagnosed in the study group (Fig 2). The 30 interval cancers in the SFM population included one case of ductal carcinoma in situ and 29 invasive cancers (tumor size was measurable in 26 invasive cancers: mean size, 24.5 mm; median size, 21.5 mm). All 10 interval cancers in the FFDM group were invasive cancers (mean size, 21.6 mm; median size, 23.5 mm).

The percentage of interval cancers for age group 50–69 years was 30 (38%) of 80 cases at SFM and 10 (24%) of 42 cases at FFDM. With the subsequent cancers in the 45–49-year age group defined as interval cancers, there were 40 interval cancers in the SFM group and 12 interval cancers in the FFDM group. The interval cancer rate was 23.6—(40/16 921) · 10 000 = 23.6)—for the SFM group and 17.4—(12/6903) · 10 000 = 17.4)—for the FFDM group (*P* = .35).

Overall True-Positive Scores and Diagnostic Performance

Twelve subsequent cancers had a true-positive score at the baseline interpretation session—nine cancers in the SFM group and three cancers in the FFDM group (Fig 3). Seven of the nine SFM cancers were dismissed at consensus meeting, and two cases had a false-negative finding at diagnostic work-up. The three FFDM cancers with a true-positive score at baseline interpretation included two cases dismissed at consensus meeting and one case with a false-negative result at diagnostic work-up.

Table 2

Screening-Detected Ductal Carcinoma in Situ Cases and Invasive Carcinomas at SFM and FFDM according to Age

Age-Group Interval (y)	No. of Women Screened	Ductal Carcinoma in Situ	Invasive	All
SFM				
45–49	7082	6 (0.08)	8 (0.11)	14 (0.20)
50–54	2531	4 (0.16)	8 (0.32)	12 (0.47)
55–59	3108	5 (0.16)	13 (0.42)	18 (0.58)
60–64	2021	1 (0.05)	5 (0.25)	6 (0.30)
65–69	2243	4 (0.18)	10 (0.45)	14 (0.62)
All (45–69)	16 985	20 (0.12)	44 (0.26)	64 (0.38)
FFDM				
45–49	2935	2 (0.07)	7 (0.24)	9 (0.31)
50–54	1051	3 (0.29)	2 (0.19)	5 (0.48)
55–59	1253	3 (0.24)	8 (0.64)	11 (0.88)
60–64	817	1 (0.12)	5 (0.61)	6 (0.73)
65–69	888	2 (0.23)	8 (0.90)	10 (1.13)
All (45–69)	6944	11 (0.16)	30 (0.43)	41 (0.59)
P value of SFM vs FFDM		.551	.040	.031

Note.—Unless otherwise indicated, data are number of cancers, with detection rate percentage in parentheses.

Figure 3

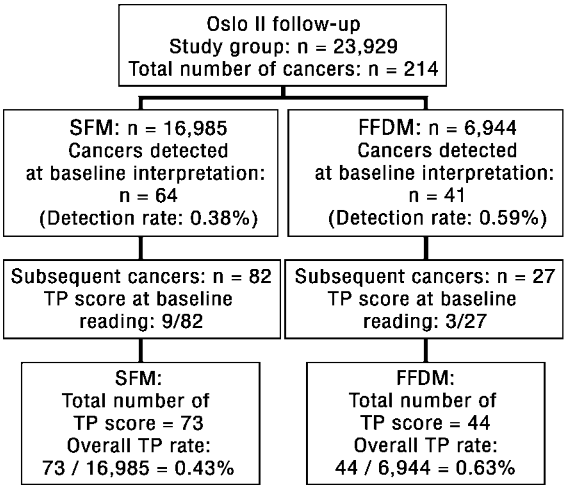

Figure 3: Flowchart shows numbers of cancers at baseline screening and subsequent cancers with true-positive scores dismissed at consensus meeting or baseline diagnostic work-up. For age group 50–69 years, the subsequent cancers included interval cancer and subsequent screening cancers. For age group 45–49 years, subsequent cancers included cancers diagnosed within 18 months after baseline screening. Overall true-positive *(TP)* score for cancers was significantly higher at FFDM than at SFM (*P* = 0.03).

Thus, the overall true-positive score was 73 (0.43%) of 16 985 cases at SFM and 44 (0.63%) of 6944 cases at FFDM. The higher true-positive score at FFDM was statistically significant (*P* = .03).

With the 12 dismissed cancers and the interval cancers grouped as false-negative findings in a 2 × 2 table analysis, sensitivity for SFM was 61.5% (64 of 104 cases) and 77.4% (41 of 53 cases) for FFDM (*P* = .07). Specificity was 97.9% (16 520 of 16 881 cases) for SFM and 96.3% (6638 of 6891 cases) for FFDM (*P* < .005) (Table 3).

Discordant Cancer Interpretations

For the 73 cancers detected at SFM with a true-positive score at baseline interpretation, both readers had a true-positive interpretation in 51 (70%) of 73 cases, whereas one of two independent readers overlooked the cancer in 22 (30%) of 73 cases (Table 4). For the 44 cancers detected at FFDM, both readers had a true-positive score in 33 (75%) of 44 cases, and only one reader had a positive score in 11 (25%) of 44 cases. When concordant interpretations were compared according to mammographic features, there were some minor differences according to lesion type, but the number of cancers in such subgroups was too small to allow us to draw conclusions.

Calculation of observer agreement for the two independent readers in the cancer cases by using quadratic weighting of the five-point rating scale showed a κ value of 0.66 for the 73 SFM cancers and a κ value of 0.55 for the 44 FFDM cancers. A confidence interval was not calculated because of a substantial number of zero cells. Linear weighting of the five-point scale revealed a κ value of 0.41 (95% confidence interval: 0.30, 0.52) for SFM and a κ value of 0.34 (95% confidence interval: 0.15, 0.52) for FFDM.

Discussion

FFDM has several potential benefits in mammographic screening and has been proposed to replace SFM in breast cancer screening. The four large-scale trials that have compared SFM and FFDM in a screening setting (1–4) and our present study show divergent results. To draw

Table 3

Performance Indicators at SFM and FFDM

Imaging Modality	No. of Women	Recall Rate (%)	Screening-Detected Cancers*	PPV_1 (%)	PPV_2 (%)	Interval Cancer Rate	Sensitivity†	Specificity†
SFM	16 985	2.5	3.8	15.1	39.0	23.6	61.5 (51.5, 70.8)	97.9 (97.8, 98.1)
FFDM	6944	4.2	5.9	13.9	42.7	17.4	77.4 (63.4, 87.3)	96.5 (96.0, 96.9)

* Number of cancers per 1000 screened women.

† Data are percentages, with 95% confidence intervals in parentheses.

Table 4

Mammographic Features and Interpretation at Independent Double Soft-Copy Reading of SFM and FFDM for all Cancers with True-Positive Scores at Baseline Reading

	SFM		FFDM	
Mammographic Feature	No. of Cancers	No. of Concordant Double Readings	No. of Cancers	No. of Concordant Double Readings
Circumscribed mass	11	6	11	8
Spiculated mass	25	21	14	12
Asymmetric density	4	0	1	0
Distortion	2	2	1	1
Microcalcifications	21	13	14	10
Density with calcifications	10	9	3	2
Total*	73	51 (70)	44	33 (75)

* Data in parentheses are percentages.

conclusions from the comparison of these studies, it is important to be aware of common aspects and differences in study design that might be the reasons for these divergent results.

The first two trials, the Colorado-Massachusetts study and the Oslo I study, may at first glance seem rather similar. Both studies had a paired design, the study groups were comparable, and in both studies the readers were inexperienced with FFDM soft-copy reading (2,5). Residents participated in some interpretations at one of the two institutions conducting the Colorado-Massachusetts study, but it seems unlikely that this had any influence on cancer detection rate. Radiologists taking part in the DMIST were all "qualified interpreters of mammograms under federal law," but there is no information about their experience in FFDM soft-copy reading (4). All radiologists participating in the Oslo II trial (the present study) had also taken part in the Oslo I study; consequently, they were all experienced in FFDM with soft-copy reading.

In the Colorado-Massachusetts study, a less powerful prototype workstation was used, and the authors suggested that an improved workstation would have possibly resulted in more cancers being detected at FFDM (5). The Oslo studies had the same mammographic unit as that in the Colorado-Massachusetts study but had a production-type workstation from the beginning. Units from different manufacturers were used in the DMIST. In the Colorado-Massachusetts study, women were offered enrollment provided that each breast was no larger than what could be depicted by standard 24 × 30-cm mammographic film (5). This exclusion criterion was not applied in the Oslo studies, and information about this aspect has not been given for the DMIST.

The Oslo I and II studies included independent double reading at both imaging modalities, whereas single reading (ie, one radiologist interpreting the SFM images and another the FFDM images) was used in the Colorado-Massachusetts study and the DMIST. Unilateral recall (ie, work-up was performed if either reader recommended it) was practiced in the DMIST (4). A filtering process (discrepancy evaluation) was included in the Colorado-Massachusetts study and the Oslo studies (consensus or arbitration meetings) before final decision for recall. There was, however, an important difference between the discrepancy evaluation in the Colorado-Massachusetts study and the consensus meeting in the Oslo studies: The discrepancy evaluation in the Colorado-Massachusetts study consisted of a side-by-side comparison of SFM and FFDM mammograms (5), whereas there were separate consensus meetings for SFM and FFDM images in the Oslo I study. Consequently, the decision for recall was made without knowledge of the results with the other imaging modality (2).

The significantly higher recall rate at FFDM (4.2%) than at SFM (2.5%) in the Oslo II study is an important difference from the Colorado-Massachusetts study results, which showed a significantly lower recall rate at FFDM (11.5%) than at SFM (13.8%) (1). Recall rates between 4.9% and 5.5% have been reported to give the best trade-off for sensitivity and PPV (9). The recall rate for SFM in our study might have been too low. Comparison images were available in the Colorado-Massachusetts study (1), but prior mammograms were not offered for the interpretation sessions in the Oslo studies to avoid interpretation bias because only prior SFM images were available. Comparison with prior examination findings significantly decreases the number of false-positive but not true-positive findings at mammographic screening (10). Of importance is that we practiced batch reading in both Oslo studies, which can help reduce recall rates without affecting the cancer detection rate (11).

Double reading can help increase cancer detection rate by 10%–15% (12–14) but may have a double impact on callback rates, depending on the recall policy used (13). We had no unilateral recall (ie, callback if either of the readers has given a positive score) in the Oslo studies because this may lead to an unacceptable number of false-positive findings. We used a low threshold for score 2 to have the possibility of a second opinion at consensus meetings. Thus, a main purpose of our consensus meetings was to increase specificity by dismissing probably benign cases. This might explain why about 70% of positive findings were dismissed at our consensus meetings compared with only 13% of positive findings at FFDM and 2.0% of positive findings at SFM at discrepancy evaluation in the Colorado-Massachusetts study (5). Although double reading by consensus or arbitration helps achieve an increase in cancer detection and a reduction in the number of women recalled for diagnostic work-up (13,15), cancers may be dismissed by using this practice. All lesions subsequently proved to be malignant may not be detected with panel consensus or arbitration (6,7,16). In nearly half of screening-detected cancers, minimal signs appeared to be present on the previous screening mammograms 2 years before diagnosis (17). The percentage of interval cancers retrospectively classified as missed constitutes 20%–35%, depending on review design (18).

Temporal instability has been shown to be an important predictive feature of malignancy among probably benign lesions (19). In the Oslo I and II studies, we did not offer prior mammograms for the reading sessions to avoid interpretation bias, and we never practice short-term follow-up. How many interval cancers might have been avoided by using another strategy remains an open question.

A total of 12 cancers—nine in the SFM group and three in the FFDM group—were dismissed at consensus meetings or had a false-negative result at work-up in our study. Dismissed findings were still counted as positive findings for calculation of recall rate, sensitivity, and other performance measures in the Colorado-Massachusetts study (1). When two diagnostic tests (SFM and FFDM) are compared, we agree that such true-positive cancers have to be included in the overall comparison and not excluded because of observer variability. The inclusion of the dismissed true-positive findings in our comparison confirmed the significantly higher cancer detection at FFDM ($P = .03$). However, these cancers are by definition not screening-detected; therefore, we excluded them when comparing the diagnostic performance by the readers for the two imaging modalities.

Diagnostic performance of the two imaging modalities was characterized by using receiver operating characteristic analysis in the DMIST, which showed a significantly higher area under the curve for FFDM in younger women and in women with dense breast parenchyma (4). There are problems with the use of receiver operating characteristic analysis in breast cancer screening (20,21), and we have therefore applied the more commonly used performance indicators (surrogate parameters) for comparison of the two modalities. Sensitivity was slightly higher for FFDM and slightly lower for SFM in our study than in the DMIST, whereas specificity was comparable when we took into account our 2-year follow-up compared with a shorter period in the DMIST. In the NBCSP, 70% of interval cancers were diagnosed in the 2nd year of the screening interval (22).

A total of 30% of the SFM cancers and 25% of the FFDM cancers in our study were given a true-positive score by only one of the two independent readers. These numbers are higher than the 13% rate of arbitration for cancers reported by other authors (14). An explanation could be that we had absolutely independent double reading because the scores were recorded directly into the database and there was no access to the database after the reading session was closed. In published studies on SFM double reading (14,16), the second reader has usually been biased by the first reader's results because of logistic reasons, and little is known about discordance in screening with truly independent double reading. Comparison of discordant interpretations according to mammographic features in our study showed a slightly higher concordance at FFDM in cancers that manifested as circumscribed masses and lower concordance in cancers that manifested as densities with calcifications. However, the numbers in each group were too small for any final conclusion. Our study was a randomized trial, and a direct comparison of discordant interpretations at SFM versus those at FFDM cannot be given as for paired studies.

In the three paired studies performed so far (1,2,4), the level of discordant interpretations at SFM versus that at FFDM for cancers has been surprisingly high. Comparison of the cancers detected with either or both modalities in a 2 × 2 table analysis shows an observed agreement of 52% and a κ value of 0.02 for the Colorado-Massachusetts study, an observed agreement of 68% and κ value of 0.37 for the Oslo I study, and an observed agreement of 66% and κ value of 0.31 for the DMIST. The slightly lower observed agreement in the Colorado-Massachusetts study could be partly explained by the use of a prototype workstation. The nearly equal κ values and observed agreements in the Oslo I study and the DMIST are noteworthy. A side-by-side feature analysis of cancers in the Oslo I study revealed that cancer conspicuity was equal for SFM and FFDM (2), and comparison of the mean scores for cancers revealed no significant difference between SFM and FFDM at independent double reading (6).

Observer agreement based on the five-point rating scale was slightly higher at SFM than at FFDM in our study, but the confidence intervals were large and the number of cancers was too small for us to derive conclusions. Reasons for discrepant interpretations of cancer have been approximately equally distributed among those relating to lesion conspicuity, lesion appearance, and interpretation (5). Thus, the large fraction of cancers missed at either of the two modalities is most likely explained by positioning variability and interpretation errors and not by failure of the imaging systems. Interobserver variability is a great challenge in mammographic screening (23,24).

Methods of evaluating technology in breast cancer screening include randomized trials with short-term follow-up and comparison of interval cancer rates (25). Reducing the rate of interval cancer is crucial, because it represents the potential benefit of early detection. The higher cancer detection rate and the lower interval cancer rate at FFDM compared with those at SFM in our Oslo II follow-up study is of interest and might indicate that FFDM is superior to SFM in mammographic screening. However, the slightly larger mean and median size of detected cancers at FFDM and the small number of cancers do not justify this suggestion, and further studies are needed to make such a conclusion.

The higher cancer detection rate at FFDM than at SFM in our Oslo II follow-up study compared with the lower detection rate in the Oslo I study is of interest. There might be several reasons for this shift to a higher cancer detection rate at FFDM. The first reason is a learning curve effect. The same radiologists participated in both Oslo studies. Although inexperienced in the Oslo I study, the readers were experienced in FFDM soft-copy reading in the Oslo II study. The second reason is the use of a dedicated screening (reading) room in the Oslo II study. Third, analysis of the false-negative FFDM interpretations in the Oslo I study directed our attention to a systematic use of the hanging protocol in batch reading (6). Fourth, analysis of the false-negative cancers at FFDM in the Oslo I study revealed a shorter reading time for many missed cancers than for normal findings (6). The mean interpretation time for normal findings at FFDM soft-copy reading in the Oslo I study was 45 seconds (6). The combination of inexperience with FFDM and too-fast soft-copy batch reading might have contributed to the greater rate of false-negative findings at FFDM in the Oslo I study.

Preliminary reports (26,27) of further studies that compare SFM and FFDM in screening settings have recently been presented at scientific meetings. A nonsignificantly lower recall rate and higher cancer detection rate at FFDM compared with those at SFM was found in the Vestfold County study (26). Results of this study showed a high detection rate for ductal carcinoma in situ, comparable to that in the DMIST and our Oslo II study. Results of the North Norway trial (27) showed a significantly higher recall rate and cancer detection rate at FFDM than at SFM, whereas the PPV was comparable for the two imaging modalities. Thus, these results are similar to those in our Oslo II study. The Vestfold County study, the North Norway study, and the two Oslo studies included the same database and logistics, but the mammographic FFDM equipment came from different manufacturers.

Limitations of our study included the interruption of the screening program for age group 45–49 years. Consequently, the follow-up was limited to 18 months for this age group. The relatively small number of cancers in each subgroup of mammographic features made comparison between SFM and FFDM at independent double reading inconclusive. Furthermore, our screening program does not record breast parenchymal density.

In conclusion, the Oslo II follow-up study results have demonstrated a significantly higher cancer detection rate at FFDM than at SFM, whereas the PPVs were comparable for the two imaging modalities. FFDM with soft-copy reading is well suited for breast cancer screening programs.

References

1. Lewin JM, Hendrick RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology 2001;218:873–880.

2. Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading—Oslo I study. Radiology 2003;229:877–884.

3. Skaane P, Skjennald A. Screen-film mammography versus full-field digital mammography with soft-copy reading: randomized trial in a population-based screening program—the Oslo II study. Radiology 2004; 232:197–204.

4. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005;353:1773–1783.

5. Lewin JM, D'Orsi CJ, Hendrick RE, et al. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. AJR Am J Roentgenol 2002;179:671–677.

6. Skaane P, Skjennald A, Young K, et al. Follow-up and final results of the Oslo I study comparing screen-film mammography and full-field digital mammography with soft-copy reading. Acta Radiol 2005;46:679–689.

7. Duijm LE, Groenewoud JH, Hendriks JH, de Koning HJ. Independent double reading of screening mammograms in the Netherlands: effect of arbitration following reader disagreements. Radiology 2004;231:564–570.

8. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Brenner RJ, Sickles EA. Analysis of 172 subtle findings on prior normal mammograms in women with breast cancer detected at follow-up screening. Radiology 2003;226:494–503.

9. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol 2001;177:543–549.

10. Burnside ES, Sickles EA, Sohlich RE, Dee KE. Differential value of comparison with previous examinations in diagnostic versus screening mammography. AJR Am J Roentgenol 2002;179:1173–1177.

11. Burnside ES, Park JM, Fine JP, Sisney GA. The use of batch reading to improve the performance of screening mammography. AJR Am J Roentgenol 2005;185:790–796.

12. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:241–244.

13. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. Breast 2001;10:455–463.

14. Cornford EJ, Evans AJ, James JJ, Burrell HC, Pinder SE, Wilson AR. The pathological and radiological features of screen-detected breast cancers diagnosed following arbitra-

tion of discordant double reading opinions. Clin Radiol 2005;60:1182–1187.

15. Mucci B, Athey G, Scarisbrick G. Double reading of screening mammograms: the use of a third reader to arbitrate on disagreements. Breast 1999;8:63–65.

16. Ciatto S, Ambrogetti D, Bonardi R, et al. Second reading of screening mammograms increases cancer detection and recall rates: results in the Florence screening programme. J Med Screen 2005;12:103–106.

17. van Dijck JA, Verbeek AL, Hendriks JH, Holland R. The current detectability of breast cancer in a mammographic screening program: a review of the previous mammograms of interval and screen-detected cancers. Cancer 1993;72:1933–1938.

18. Hofvind S, Skaane P, Vitak B, et al. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology 2005;237:437–443.

19. Rosen EL, Baker JA, Soo MS. Malignant lesions initially subjected to short-term mammographic follow-up. Radiology 2002;223:221–228.

20. Keen JD. Digital and film mammography [letter]. N Engl J Med 2006;354:765–767; author reply 765–767.

21. Skaane P, Niklason L. Receiver operating characteristic analysis: a proper measurement for performance in breast cancer screening? AJR Am J Roentgenol 2006;186:579–580.

22. Hofvind S, Bjurstam N, Sørum R, Bjørndal H, Thoresen S, Skaane P. Number and characteristics of breast cancer cases diagnosed in four periods in the screening interval of a biennial population-based screening programme. J Med Screen 2006;13:192–196.

23. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–1499.

24. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996;156:209–213.

25. Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer. BMJ 2006;332:678–679.

26. Vigeland E, Hofvind SS, Klaasen H, Wegener A, Abrahamsen A, Skaane P. Population-based screening using full field digital mammography (FFDM) with soft copy reading. First year experience from Vestfold County, Norway [abstr]. In: Radiological Society of North America scientific assembly and annual meeting program. Oak Brook, Ill: Radiological Society of North America, 2005; 287.

27. Bjurstam N, Frantzen JO, Pedersen K, Hofvind S. Full-field digital mammography screening in the population-based screening program in North-Norway: preliminary results [abstr]. In: Radiological Society of North America scientific assembly and annual meeting program. Oak Brook, Ill: Radiological Society of North America, 2006; 392.