Eur Radiol (2013) 23:2061–2071
DOI 10.1007/s00330-013-2820-3

BREAST

Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration

Per Skaane · Andriy I. Bandos · Randi Gullien · Ellen B. Eben · Ulrika Ekseth · Unni Haakenaasen · Mina Izadi · Ingvild N. Jebsen · Gunnar Jahr · Mona Krager · Solveig Hofvind

Received: 12 December 2012 / Revised: 31 January 2013 / Accepted: 2 February 2013 / Published online: 4 April 2013

Abstract
Objectives To compare double readings when interpreting full field digital mammography (2D) and tomosynthesis (3D) during mammographic screening.
Methods A prospective, Ethical Committee approved screening study is underway. During the first year 12,621 consenting women underwent both 2D and 3D imaging. Each examination was independently interpreted by four radiologists under four reading modes: Arm A—2D; Arm B—2D+CAD; Arm C—2D+3D; Arm D—synthesised 2D+3D. Examinations with a positive score by at least one reader were discussed at an arbitration meeting before a final management decision. Paired double reading of 2D (Arm A+B) and 2D+3D (Arm C+D) were analysed. Performance measures were compared using generalised linear mixed models, accounting for inter-reader performance heterogeneity ($P<0.05$).
Results Pre-arbitration false-positive scores were 10.3 % (1,286/12,501) and 8.5 % (1,057/12,501) for 2D and 2D+3D, respectively ($P<0.001$). Recall rates were 2.9 % (365/12,621) and 3.7 % (463/12,621), respectively ($P=0.005$). Cancer detection was 7.1 (90/12,621) and 9.4 (119/12,621) per 1,000 examinations, respectively (30 % increase, $P<0.001$); positive predictive values (detected cancer patients per 100 recalls) were 24.7 % and 25.5 %, respectively ($P=0.97$). Using 2D+3D, double-reading radiologists detected 27 additional invasive cancers ($P<0.001$).
Conclusion Double reading of 2D+3D significantly improves the cancer detection rate in mammography screening.
Key Points
- *Tomosynthesis-based screening was successfully implemented in a large prospective screening trial.*
- *Double reading of tomosynthesis-based examinations significantly reduced false-positive interpretations.*
- *Double reading of tomosynthesis significantly increased the detection of invasive cancers.*

P. Skaane
Department of Radiology, Oslo University Hospital, University of Oslo, Oslo, Norway

A. I. Bandos
Department of Biostatistics, University of Pittsburgh, Pittsburgh PA, USA

R. Gullien · E. B. Eben · U. Haakenaasen · M. Izadi · I. N. Jebsen · G. Jahr · M. Krager
Department of Radiology, Oslo University Hospital, Oslo, Norway

U. Ekseth
Curato Roentgen Institute, Oslo, Norway

S. Hofvind
Institute of Population-based Cancer Research, The Cancer Registry, Oslo, Norway

P. Skaane (✉)
Department of Radiology, Breast Imaging Center, Oslo University Hospital Ullevaal, Kirkeveien 166,
0407 Oslo, Norway
e-mail: PERSKA@ous-hf.no

Keywords Breast cancer screening · Mammography · Double reading · Full-field digital mammography · Tomosynthesis

Abbreviations

FFDM (2D)	Full-field digital mammography
2D+3D	Full-field digital mammography plus tomosynthesis
CAD	Computer-aided detection
DBT	Digital breast tomosynthesis
FDA	US Food and Drug Administration
NBCSP	Norwegian Breast Cancer Screening Program
OTST	Oslo Tomosynthesis Screening Trial

Introduction

Periodic mammographic screening has been found to result in the earlier detection of breast cancers, leading to a reduction in patient mortality and morbidity [1–3]. Screen-film mammography (SFM) was the standard technique in breast cancer screening for many years, but today the most common imaging procedure is a two-view (medio-lateral oblique and cranio-caudal) examination using full-field digital mammography (FFDM). The success of screening mammography depends on the detection of small, subtle non-palpable cancers which may be a very difficult task. Consequently, inter-observer and intra-observer variability that may be affected by many factors such as case difficulty, radiologist's experience, varying practices and others, is a great challenge in mammographic screening and has been shown to be a great problem for SFM as well as FFDM [4, 5]. In a nationwide mammography screening programme using independent double reading, a total of 23 % of the screening-detected cancers had a discordant interpretation, i.e. a true-positive score by only one of the two readers [6]. The differences in mammography interpretation can influence cancer detection and consequently the effect of screening mammography. Double reading of screening mammograms has therefore been recommended as a measure to increase the cancer detection rate. Double reading without further management decisions would, however, substantially increase the recall rate. Several European organised mammography screening programmes have therefore implemented double reading with a consensus or arbitration procedure before making the final decision whether to recall or not recall women who have had an examination with positive scores by one or both readers. Although this approach requires substantial professional resources both during the double-reading step and during the consensus/arbitration meeting, this method provides a high cancer detection rate while maintaining a low recall rate [7–16].

Recently, tomosynthesis-based imaging procedures have been implemented for the purposes of screening for the early detection of breast cancer. The principles of the approach have been described elsewhere [17]. In brief, a series of low-dose projection images (2D) are acquired at different angles along an arc and using a filtered back-projection reconstruction method. The multi-view information from the multiple low-dose images is used to generate thin slices (at 1-mm spacing) that can be viewed sequentially as a stack. The primary operational advantage of tomosynthesis-based imaging is that the procedure is very similar to a conventional FFDM-based examination in terms of the technologist's tasks and the woman being imaged; therefore, tomosynthesis can be easily implemented in current clinical practices with minor operational adjustments. There are preliminary indications that the use of tomosynthesis with a single interpreter increases cancer detection, while at the same time decreasing recall rates [18]. In addition, a number of retrospective and experimental clinical studies evaluating tomosynthesis, primarily in laboratory settings using cancer-enriched populations, demonstrated the potential for decreasing recall rates and possibly increasing cancer detection rates [19–29]. While these preliminary results suggest that an increase in cancer detection rate and a simultaneous decrease in recall rate is achievable using tomosynthesis, none of these studies assessed the performance of tomosynthesis when double reading is employed in daily practice. Hence, to our knowledge there are no data to date on the impact of double reading in a tomosynthesis-based screening environment.

The large prospective single institution Oslo Tomosynthesis Screening Trial (OTST), part of a population-based mammography screening programme, has four study arms interpreted independently by four radiologists: conventional FFDM (2D), FFDM plus computer-aided detection (2D+CAD), conventional FFDM plus tomosynthesis (2D+3D), and synthesised 2D plus tomosynthesis (synthesised 2D +3D). The design of the trial includes two very similar 2D-alone-based arms and two very similar 2D+3D-based arms allowing for a comparison between independent double reading of 2D and independent double reading of 2D+3D. The purpose of our study was to analyse the performance in terms of cancer detection, false-positive scores before arbitration, and actual recall rates among the 12,621

participants who were imaged during the first year of the OTST.

Materials and methods

Study population

The trial was approved by the Ethics Committee with written informed consent required by all participants. All women included in this study were invited by a personal letter to participate in the Breast Cancer Screening Program in Oslo between 22 November 2010 and 31 December 2011. The Oslo screening program is part of the Norwegian Breast Cancer Screening Program (NBCSP) administered by the Cancer Registry of Norway. This screening programme, inviting women aged 50–69 years to two-view mammography biennially, has been described in detail elsewhere [30, 31].

Upon arrival for the scheduled examination women were asked if they were willing to participate in the study. The selection of potential candidates to be asked to participate was based solely on the availability of technical staffing on the date of the examination in question and the availability of imaging systems to perform the additional imaging procedures, and not based on any personal information about the women who were approached to consider participation. Disabled women (e.g. those unable to stand) and women with breast implants were not asked to consider participation (excluded). We were not permitted by the ethics committee to record the reason for declining to participate in the trial. Women who were not asked to participate (due to the unavailability of staffing and/or imaging systems) and women who were asked but declined to participate underwent conventional FFDM imaging.

Independent double reading is standard practice in our screening programme and the assessment of tomosynthesis in this prospective clinical trial was incorporated into our routine biennial screening mammography practice. We provide here the results of independent double reading of 2D alone versus 2D+3D ("combo mode") of examinations that were acquired from consenting women during the study period and as a result underwent interpretation in all four arms.

Imaging technique

All screening examinations were carried out in the screening unit located in downtown Oslo. The screening unit was equipped with three commercially available imaging systems that included the capability to obtain examinations using FFDM as well as tomosynthesis (Dimensions; Hologic, Bedford, MA, USA). After image quality assurance was given by the technologist, all imaging examinations were transferred to the Breast Imaging Center at the Oslo Hospital, Ullevaal for interpretation and management recommendation.

Conventional FFDM (2D) as well as tomosynthesis included a two-view mammography imaging procedure (cranio-caudal and medio-lateral-oblique) of each breast. Both the conventional FFDM and the 3D-tomosynthesis imaging were acquired during a single breast compression per view. This combined imaging time of the procedure took approximately 10 s per view. The radiation dose of the tomosynthesis imaging was automatically determined by the imaging system and was set to result in approximately the same dose as a single mammographic view.

Reading modes and the generation of synthesised 2D images

A general description of the four-arm prospective study with analysis of a single reader approach for FFDM alone and FFDM+tomosynthesis has been presented elsewhere [18]. The four interpretation modes included: (1) Arm A: 2D only; (2) Arm B: 2D+computer-aided detection (CAD); (3) Arm C: 2D+3D; (4) Arm D: a synthesised 2D+3D (in which synthesised 2D images were reconstructed from the 3D dataset and hence did not require additional exposure of the breast). The CAD system used for Arm B was a commercially available system (ImageChecker 9.3; Hologic).

The synthesised images used in this analysis were reconstructed using an early version of image reconstruction. An improved version of the one used here was recently reviewed by the FDA for possible pre-market approval [32]. The synthesised 2D image is created by summing and filtering the stack of reconstructed tomosynthesis slices similar to generating a maximum intensity projection (MIP) image. This image-processing approach was developed by Hologic, and a more detailed description of the method is described elsewhere [33].

The acquisition protocol resulted in fully registered 2D and 3D images as the 2D and 3D images for each view were obtained under a single compression.

Reader training

Before commencement of the trial, all radiographers and radiologists participating in the trial received specific training in the operation of the imaging system and in the interpretation of tomosynthesis examinations [18, 28]. Eight radiologists with 2–31 years of experience (average 16 years) in screening mammography participated in this

study. Each of the eight participating radiologists received individualised intensive personal training of approximately 4 h in reviewing an enriched set of a minimum of 100 examinations with feedback, using the same workstations used in the trial and the respective hanging protocols.

Image hanging protocols and workflow

The women invited to our screening programme, as part of the NBCSP, were scheduled for a two-view FFDM examination with independent double reading according to the Norwegian guidelines. Consequently, the women had to be offered the 2D examinations with double reading (Arm A and Arm B) provided. Our study design also included independent double reading for the two tomosynthesis arms (Arm C and Arm D). Thus, each examination included in the trial was independently interpreted by four different radiologists in a batch mode using four dedicated workstations, one workstation for each arm.

Reading assignments were made by a non-radiologist staff member who independently assigned each radiologist sets of cases to be interpreted under a specific reading mode. The scheduler attempted to balance the number of cases interpreted by each radiologist under each mode as much as was reasonably achievable in a busy clinical practice in which some of the participating radiologists are not present full time in the Breast Imaging Center. During the batch readings, hanging protocols for the specifically assigned modes were pre-set. Previous digital screening mammograms were reviewed when available, but previous screen-film mammograms from our hospital or other institutions, if available, were only reviewed at the consensus meeting if a positive score had been given in one of the four arms.

The first few steps of the hanging protocol were common for all four arms. First, all four views were displayed: the two cranio-caudal views back-to-back on the left monitor and the two medio-lateral oblique views back-to-back on the right monitor. If previous digital screening mammograms were available for comparison, these were displayed at the top and the current images at the bottom. In the second step, both current cranio-caudal views were presented at "full size", one on each monitor. Next, both medio-lateral oblique views were presented at "full size", one on each monitor. The last step for Arms A and B included all views once again for comparison (similar to the first step). The further steps for the 2D+3D interpretations (Arms C and D) then included the 2D image for each view displayed on the left monitor and the 3D image for the same view was displayed on the right monitor, the cranio-caudal views first followed by the medio-lateral oblique views. The last step for the two "combo arms" included the first step once again, similar to Arms A and B.

Image interpretation and consensus/arbitration decision

As this is a prospective clinical trial incorporated into routine clinical practice, the reading environment was not precisely monitored. The same light-dimmed reading rooms used in our clinical practice were used in the trial. After reviewing an examination, each radiologist independently rated his/her findings per breast using a standardised five-point ordinal rating scale [18]. This five-point rating scale for probability of cancer has the following classifications: 1=normal or definitely benign; 2=probably benign; 3=indeterminate; 4=probably malignant; 5=malignant. A score of 1 by all four readers is regarded as a negative examination. The decision to undertake additional actions other than dismissal of a case as negative was based solely on the ratings of any of the four radiologists (one or more) recording a score of 2 or higher (≥2). In these instances, mammographic findings (features) had to be listed as well. Other indications for selection for the consensus meeting included the presence of clinical symptoms, especially a palpable lump, or technical insufficiency of the examination. Scores (ratings) were recorded directly into the NBCSP database, and the results were locked at the end of each reading session.

All cases receiving one or more scores of 2 or greater in at least one reading arm were discussed at arbitration, with at least two radiologists participating in these meetings and with availability of all imaging and non-imaging information. A consensus-based clinical management decision (dismiss or recall for diagnostic work-up) was reached for all examinations receiving at least one rating of 2 or 3. An examination receiving a score of 4 or 5 was recalled and could not be dismissed at consensus. The breast parenchyma density was assessed in consensus according to the American College of Radiology (ACR) in four categories and recorded. Diagnostic work-up of recalled women that could potentially include additional views, ultrasound, magnetic resonance imaging (MRI) and needle biopsy if indicated, was performed during a single visit to the Breast Imaging Center by the same group of radiologists.

Furthermore, the interpretation time (in seconds) for each reader in each of the four arms was automatically recorded directly into the NBCSP database. The time registration started automatically when the score sheet was fetched to the screen using bar code technology, and the time registration was stopped when the reader clicked on "save" using the mouse. As examinations were already uploaded onto the designated workstation

and uploading images onto the display of the workstation is almost instantaneous (approximately a second), the measured reading-time reasonably adequately represents the actual time the radiologist took to view and interpret each case.

In our breast cancer screening programme, short-term (6-month) follow-up is not recommended, neither at arbitration meeting nor after recall with diagnostic work-up. Women with a positive score are either dismissed and scheduled for the next screening round after 2 years or a complete work-up including needle biopsy is performed, if indicated.

Statistical analyses

As independent interpretation with exact duplication of reading conditions for multiple radiologists are not implemented in the trial, for the purposes of our analyses we assume that the paired arms of the study for 2D alone (Arm A plus B) and 2D+3D (Arm C plus D) constitute sufficiently similar reading conditions to be considered double readings. Therefore, we combined the positive scores from either of the paired 2D or the paired 2D+3D reading modes. All analyses were performed considering a significance level of 0.05. Inferences about relative changes of the rates adjusted for radiologists' performances and correlations between assessments of the same cases were conducted using a Type III test in the context of a generalised linear mixed model (proc glimmix, v. 9.23; SAS, Cary, NC, USA). Heterogeneity of the performance levels of different combinations of radiologists under 2D and 2D+3D reading modes were addressed using G-side random effects.

We compared false-positive rates, attributable recall rates as a result of arbitration decisions, attributable cancer detection rates and positive predictive values (verified attributable cancers/recalls) for paired independent double readings of 2D only and 2D+3D using the following outcome measure assignments. A screening examination with a positive score (i.e. receiving a score of 2 or higher by at least one of the two readers during the initial interpretation under either of the two modes in question) that was later confirmed to have cancer as a result of diagnostic work-up, was attributed as "detected" (true positive) under the specific double reading method, namely, 2D or 2D+3D, respectively. A case with a positive score without a verified cancer (either dismissed at the arbitration meeting or determined as benign during the diagnostic work-up, which may have included a biopsy) was considered a false positive to that double-reading method. Positive predictive value (PPV) was computed as the percentage of cases with screen-detected cancer among all positively scored cases (positive score for at least one breast) that were recalled at the arbitration meeting. Predictive values were compared in the setting of the generalised linear model, adjusting for the heterogeneity of the performance levels of different combinations of radiologists evaluating different images. Odds ratio (PPV odds for 2D relative to PPV odds for 2D+3D) and the corresponding 95 % confidence interval were computed and compared.

Results

Study population

Between November 22, 2010 and December 31, 2011, a total of 29,652 women were invited to the screening programme. Attendance during this period was 17,960 or 60.6 %. A total of 5,329 women underwent 2D mammography only (Arms A and B). Among these women, 31 cancers were diagnosed (cancer detection rate 0.58 %). Women having only 2D were excluded from further analysis. Thus, we recruited 12,631 women (70.3 % of all women attending the screening programme) who consented to participate in our tomosynthesis assessment trial during the study period (Fig. 1). After excluding ten patients from analysis (three interval cancers, five with metastases or malignant lymphomas and two women with palpable cancer and normal mammographic scores by all four readers), the remaining 12,621 represent our study population (Fig. 1). The age range of the studied population was 50–69 years old (average 59.3 years old).

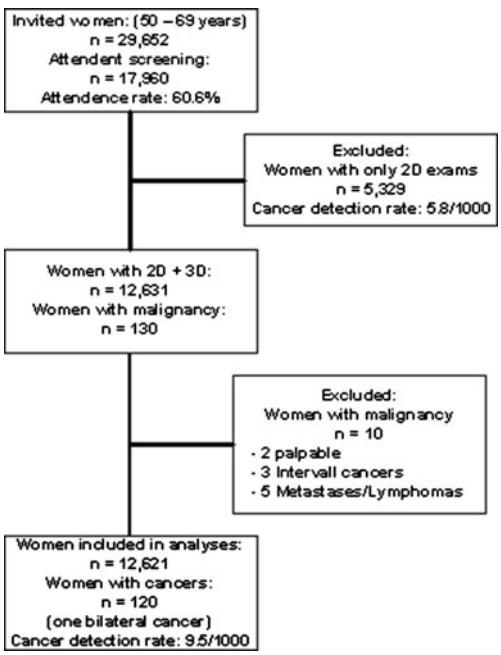

Fig. 1 Flow chart showing the number of women attending the screening programme during the study period, the number of women excluded from analysis and the study material

Positive interpretation scores

We note that "positive" scores represent a pre-arbitration measure of suspicion level by one radiologist and are not actual recall rates. From the 2D-based double-reading mode, a total of 1,382 cases (848 and 850 from Arms A and B, respectively; note some cases were given a positive score by both Arms A and B) with positive scores were identified compared with 1,175 cases with positive scores (771 and 676 from Arms C and D, respectively) from the 2D+3D double-reading mode. These results largely confirm the assumption that the modes used for "double-reading analyses" were reasonable and adequate for the purpose of the estimates presented in this paper. From these referrals, 120 patients were later found to have cancer (1 bilateral). As a result, the false-positive rates were 10.3 % (1,286/12,501) and 8.5 % (1,057/12,501) for the 2D alone and 2D+3D modes, respectively (18 % decrease, $P<0.001$). Six women with breast cancer in the contra-lateral breast of that scored as positive were not considered as false positives in the analysis. Of women with positive scores for 2D and 2D+3D double-reading modes, 365 and 463 respectively were recalled at the arbitration ($P=0.005$). These patient-level results are shown in Fig. 2.

Cancer detection

The diagnostic work-up of 540 patients recalled after arbitration (total patients recalled from the 2D arm, 2D+3D arm or both arms) resulted in the detection of 90 cancers (from 90 patients) under the 2D alone double reading mode and 119 cancers (from 118 patients, 1 had screen-detected bilateral cancer) under the 2D+3D double-reading mode, respectively. The cancer detection rate for 2D double reading was 0.71 % (or 7.1 cancers/1,000 women screened) compared with 0.94 % (or 9.4 cancers/1,000 women screened) for the 2D+3D double-reading mode ($P<0.001$), which represents a reader-adjusted increase in cancer detection of 30 % (Fig. 2). A summary of the detected cancers is provided in Table 1. There were two cancers that were detected by the 2D readers only and 31 cancers that were detected by the 2D+3D mode readers only. Overall, the proportion of detected cancers are 0.74 (90/121) and 0.98 (119/121) for 2D and 2D+3D double readings, respectively.

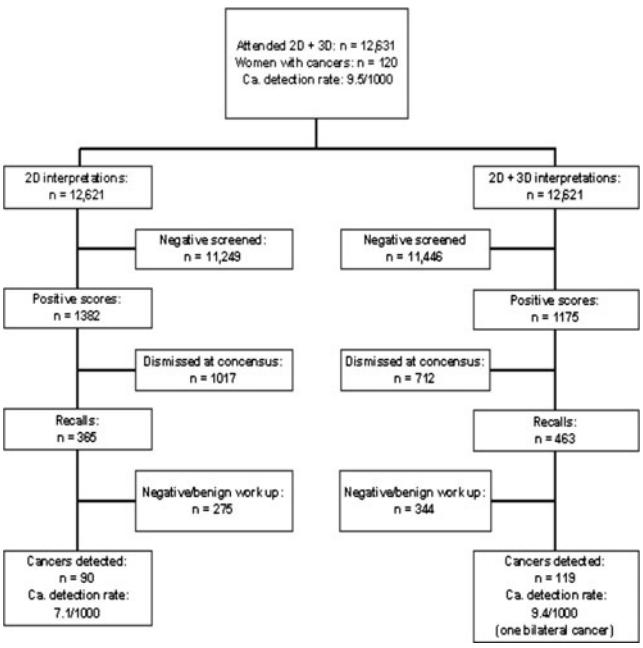

Fig. 2 Flow chart showing the positive scores, the recalls and the cancers detected in the 2D arm and the combined 2D+3D (combo mode) for the study population

The PPV for detected cancers per recall that were attributed to the two double-reading modes were 24.7 % (90/365) and 25.5 % (118/463), respectively (reader-adjusted odds ratio of 0.99 with 95 % CI from 0.69 to 1.42). Twenty-four of the 29 additional cancers detected under the 2D+3D mode were node-negative invasive cancers, 21 of which were depicted as spiculated masses and/or distortions. A summary of the cancers is provided in Table 1. Figures 3 and 4 demonstrate a cancer missed by both 2D readers and detected by the 2D+3D readers. As can be seen in Table 1, additional cancers were detected by the 2D+3D readers in all breast density categories.

Average interpretation times were 48 and 89 s per reading for the 2D and 2D+3D modes, respectively ($P<0.001$). Average system computed mean fibro-glandular doses for the 2D (mode A or B), the 3D plus synthetic 2D (mode D) and the 2D+3D (mode C) were 1.58 ± 0.61 mGy, 1.95±0.58 mGy, and 3.52±1.08 mGy, respectively.

Discussion

To the best of our knowledge, this is the first analysis comparing the performance of double reading of FFDM with a double reading of tomosynthesis (2D+3D) in a large prospective clinical trial. Several previous studies, mainly in experimental clinical settings, have shown that tomosynthesis has the potential to detect more cancers than conventional 2D mammography [26, 28]. The results of this study, demonstrating a relative increase in cancer detection of 30 %, shows that tomosynthesis may play a major potential role in mammography screening. Perhaps most important in terms of detection is the fact that most additional cancers detected using tomosynthesis tend to be invasive, with a large fraction being node-negative. In this double-reading analysis, which is likely to be applicable to a large number of clinical practices, we found similar

Table 1 Distribution of detected cancers by different categories and features

			Detected cancer			
			2D only	2D+3D only	2D and 2D+3D	All
All	Number		2	31	88	121[a]
Invasive (± DCIS)	Number		2	29	65	96
	Histology	IDC	0	15	42	57
		IDC+DCIS	0	7	12	19
		ILC	2	7	8	17
		Other primary invasive cancers	0	0	3	3
	Lymph node status	Negative	2	26	48	76
		Positive	0	2	13	15
		Unknown status	0	1	4	5
	Grade	1	1	14	22	37
		2	1	11	32	44
		3	0	3	10	13
		Unknown grade	0	1	1	2
	Breast density	1 – Fatty	0	2	4	6
		2 – Scattered densities	0	11	33	44
		3 – Heterogeneously dense	2	13	25	40
		4 – Extremely dense	0	3	3	6
	Radiological sign	Calcifications	0	0	6	6
		Mass and calcifications	0	6	6	12
		Circumscribed mass	0	0	9	9
		Spiculated mass	1	10	32	43
		Architectural distortion	0	11	9	20
		Asymmetric density	1	2	3	6
	Size (mm)	≤10	1	15	29	45
		11–15	0	12	15	27
		16–19	1	0	5	6
		≥20	0	1	14	15
		No size	0	1	2	3
		Mean	14	11	14	13
		Median	14	10	13	11
		Minimum	9	5	1	1
		Maximum	19	22	50	50
DCIS	Number		0	2	23	25
	Grade	Low/medium grade	0	0	4	4
		High grade	0	2	18	20
		Missing	0	0	1	1

DCIS ductal carcinoma in situ

[a] 121 cancers were detected in 120 women (1 woman with bilateral cancer)

results to those obtained in a single-reading environment [18]. As in the single-reader analyses, we did not find a substantial number of additional ductal carcinomas in situ (DCISs) being detected.

Previous studies in a single-reader environment have shown a reduction in false-positive scores [18, 21]. Our results for double reading are in agreement with these findings. As in the previous study [18], the recall rate after the arbitration meeting for the 2D+3D double reading was higher than with 2D only (463 versus 365, respectively). However, the 2D+3D double reading approach resulted in the detection of 29 additional cancers. Although 98 additional women were recalled using the 2D+3D reading mode compared with 2D alone, the

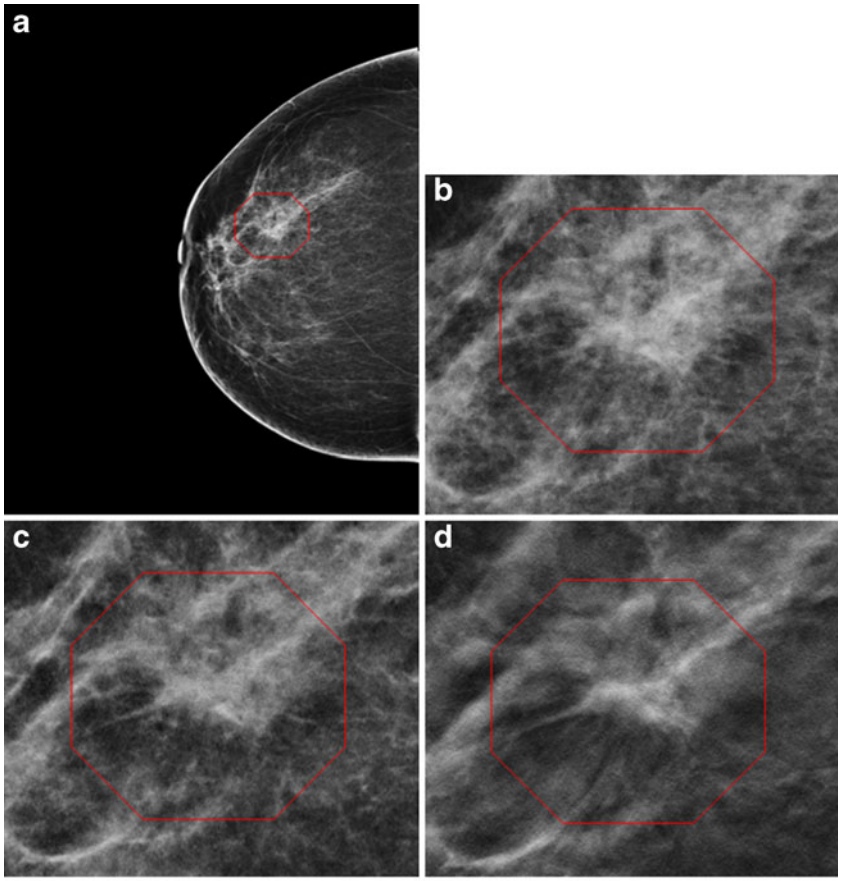

Fig. 3 A 68-year-old woman. **a** Right breast cranio-caudal view (2D) shows a non-specific density. Enlargement of the 2D (**b**) and synthesised 2D (**c**) shows a suspicious but non-conclusive irregular density. On tomosynthesis (3D) cranio-caudal view, however, a spiculated mass consistent with invasive cancer is clearly seen (**d**). The cancer was missed by both readers in the 2D arm. Histology revealed an 8-mm invasive lobular carcinoma grade 2

cancer detection "positive predictive value" (PPV) for these additional recalls was 30 % (29/98). Overall the PPV was similar for the two reading modes (24.7 % and 25.5 %, respectively) with most additionally detected cancers under the 2D+3D mode being invasive, node-negative or the very types of cancers one wishes to find during screening (Table 1). As the study was designed for multiple (four) interpretations of the same examination by four different radiologists, our results showed a higher recall rate for 2D+3D. The higher recall rate for tomosynthesis than for the 2D only mode is not in agreement with results from previous studies not employing an arbitration step in the decision process and is most likely explained by bias in favour of the 2D mode at the arbitration meetings. This likely bias in favour of the 2D only method results from the fact that several unrecalled cases with a positive score with the 2D only mode but with normal findings with the tomosynthesis mode would possibly have been recalled at an arbitration meeting if 3D had not been available. In support of this statement we observed that, compared with the average recall rate of 4.1 % during the last FFDM-only screening cycle, the recall rate was 2.9 % when tomosynthesis was available during arbitration of examinations referred by one or both FFDM-only interpreters. This difference (~1.2 %) alone would have resulted in approximately 150 additional recalls from the FFDM only double reading arm in this study.

We used synthesised 2D images in one of the two 2D+3D arms in our study. While current true 2D+3D requires more than doubling (2.2 times) of the radiation dose to the breast being imaged, if indeed synthesised images are used for this purpose, the radiation dose can be reduced substantially to comparable levels to those used in 2D imaging (in our study 1.2 times the dose for FFDM alone) with significantly improved performance over the 2D-only double reading approach [34]. The image processing used is designed to generate synthesised 2D images that "look" like conventional full-field digital mammograms enabling the radiologist to use the synthetic 2D image during the interpretation as he or she would a conventional FFDM image, namely, for comparison with previous studies, identification of mass-like abnormalities and/or distortions, assessment of left–right breast asymmetry, and detection of microcal cification clusters. The primary interest in these synthesised images as related to this work lies in the fact that reconstructing the synthetic 2D images from the 3D data sets does not require any additional radiation exposure to the breast being imaged.

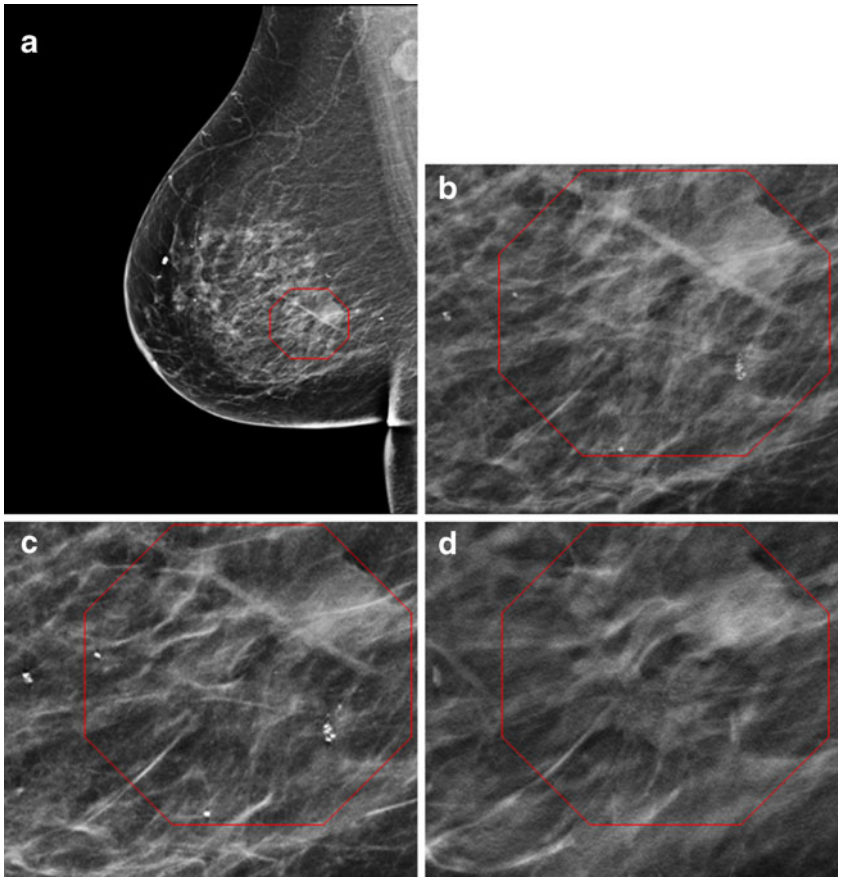

Fig. 4 A 64-year-old woman. **a** Right breast medio-lateral-oblique view (2D) shows a round benign mass shown in the upper right portion of the marked region, but no suspicious findings. Enlargement of 2D (**b**) and synthesised 2D (**c**) does not show any suspicious findings. Tomosynthesis (3D) (**d**) medio-lateral-oblique view clearly shows a spiculated mass ventro-caudal of the benign round mass and in the centre of the marked region. The cancer was missed by both readers in the 2D arm. Histological examination showed a 6-mm tubular carcinoma

The optimal method of using tomosynthesis in mammography screening needs to be addressed. In an experimental clinical setting, the performance of tomosy nthesis using only one view was not inferior to 2D digital mammography using two views [20]. Also, the combination of single-view digital breast tomosynthesis (DBT) with the opposite 2D FFDM view yields significantly superior diagnostic accuracy compared with dual-view FFDM [23, 35]. Another study reported improvements in observer performance levels for the combined reading mode compared with FFDM alone [36]. Based on our experience in using tomosynthesis in a clinical setting, we decided to use the combination of 2D+tomosynthesis ("combination mode") in both views (cranio-caudal as well as medio-lateral-oblique) in our screening trial.

Reading time for tomosynthesis will be an important and crucial aspect if this new technology is going to be implemented in organised, population-based screening programmes. The time to interpretation is, of course, longer for the combined 2D+3D mode than for 2D alone [37]. There was, however, considerable variation among the eight radiologists in this study. A learning curve effect and future improvements in hanging protocols might reduce reading times. Although there was almost a doubling of the interpretation time for the combined 2D+3D mode compared with 2D alone, we think that the increase from 48 s to 89 s is acceptable, taking into account the substantial increase in cancer detection of more than 30 % demonstrated in this study.

Our study has several limitations. Firstly, the analysis we performed simulates closely what one would expect under a double reading environment, but it was not a pure double reading experiment. Each examination was independently read by four radiologists who could refer cases to arbitration. In addition, the 2D+3D double-reading mode included one reading of actually acquired FFDM+3D, while the other mode included synthesised 2D images + 3D. Similarly, the two 2D-only reading modes were not identical in that CAD was available in one of the reading modes. However, our results suggest that the effect, if any, of these non-identical modes within each of the double-reading approaches is small as the paired modes being considered as "double reading" (i.e. Arms A and B or C and D) were quite similar in terms of positive scores. The pairs (Arms A and B versus C and D) differed between them significantly less than the difference between the modes without and with tomosynthesis. Secondly, the consensus/arbitration step could have preferentially decreased actual recall rates of women referred to arbitration (rating ≥2) under only one of the modes that were

later dismissed during arbitration; however, we found no evidence to date that interval cancers would have belonged to this group or would have changed any of our conclusions. Thirdly, despite substantial effort, we could not completely balance the interpretation load (reader by mode)—having each reader interpret the same number of examinations under each of the reading modes—as this turned out to be a very difficult task in a very busy clinical environment with some of the radiologists not working full time at the clinic. We accounted for this imbalance by adjusting for differences in method-specific performance levels of the combinations of the radiologists interpreting specific images. Despite these limitations, we believe the results of this study are valid, in particular in terms of relative performance differences, and we anticipate that similar effects would likely be observed in a true double-reading environment.

In conclusion, we found that double reading using tomosynthesis-based imaging resulted in a significant increase in cancer detection rates, specifically in the detection of invasive, node-negative cancers, and at the same time a reduction in false-positive scores compared with a double reading of 2D imaging alone.

Acknowledgements The University of Oslo (Dr. Skaane, PI) received support from Hologic in the form of equipment on loan and financial support for the additional readings required in this project. The University of Pittsburgh (Dr. Bandos, PI) received a contract to independently analyse the results of this study. We have previously reported the first-year results for two of four study arms (single-reader comparison of 2D compared with 2D+3D) [18], which compared interpretation using the North American model of single reading. Our current submission is an analysis including the interpretations from all four arms of the study using the European standard of independent double reading.

References

1. Paap E, Holland R, den Heeten GJ et al (2010) A remarkable reduction of breast cancer deaths in screened versus unscreened women: a case-referent study. Cancer Causes Control 21:1569–1573
2. Tabar L, Vitak B, Chen THH et al (2011) Swedish two-county trial: Impact of mammographic screening on breast cancer mortality during 3 decades. Radiology 260:658–663
3. EUROSCREEN Working Group (2012) Summary of the evidence of breast cancer service screening outcomes in Europe and first estimate of the benefit and harm balance sheet. J Med Screen 19 (Suppl 1):5–13
4. Beam CA, Layde PM, Sullivan DC (1996) Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 156:209–213
5. Skaane P, Diekmann F, Balleyguier C et al (2008) Observer variability in screen-film mammography versus full-field digital mammography with soft-copy reading. Eur Radiol 18:1134–1143
6. Hofvind S, Geller BM, Rosenberg R et al (2009) Screening-detected breast cancers: Discordant independent double reading in a population-based screening program. Radiology 253:652–660
7. Duijm LEM, Groenewoud JH, Hendriks JHCL, de Koning HJ (2004) Independent double reading of screening mammograms in the Netherlands: effect of arbitration following reader disagreement. Radiology 231:564–570
8. Shaw CM, Flanagan FL, Fenlon HM, McNicholas MM (2009) Consensus review of discordant findings maximizes cancer detection rate in double-reader screening mammography: Irish national breast screening program experience. Radiology 250:354–362
9. Ciatto S, Ambrogetti D, Bonardi R et al (2005) Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening program. J Med Screen 12:103–106
10. Swensson RG, King JL, Good WF, Gur D (2000) Observer variation and the performance accuracy gained by averaging ratings of abnormality. Med Phys 27:1920–1933
11. Metz CE, Shen JH (1992) Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. Med Decision Making 12:60–75
12. Harvey SC, Geller B, Oppenheimer RG et al (2003) Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 180:1461–1467
13. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J (2001) Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. Breast 10:455–463
14. Ciatto S, Ambrogetti D, Risso G et al (2005) The role of arbitration of discordant reports at double reading of screening mammograms. J Med Screen 12:125–127
15. Cornford EJ, Evans AJ, James JJ, Burrell HC, Pinder SE, Wilson ARM (2005) The pathological and radiological features of screen-detected breast cancers diagnosed following arbitration of discordant double reading opinions. Clin Radiol 60:1182–1187
16. Matcham NJ, Ridley NTF, Taylor SJ, Cook JL, Scolding J (2004) Breast screening: the use of consensus opinion for all recalls. Breast 13:184–187
17. Ren B, Ruth C, Wu T et al (2010) A new generation FFDM/tomosynthesis fusion system with selenium detector. Proc SPIE 7622:B1–B10
18. Skaane P, Bandos AI, Gullien R, et al. Comparing digital mammography alone versus digital mammography plus tomosynthesis in a population-based screening program. Radiology. doi: 10.1148/radiol.12121373
19. Bernardi D, Ciatto S, Pellegrini M et al (2012) Prospective study of breast tomosynthesis as a triage to assessment in screening. Breast Cancer Res Treat 133:267–271
20. Gennaro G, Toledano A, di Maggio C et al (2010) Digital breast tomosynthesis versus digital mammography: a clinical performance study. Eur Radiol 20:1545–1553
21. Gur D, Abrams GS, Chough DM et al (2009) Digital breast tomosynthesis: Observer performance study. AJR Am J Roentgenol 193:586–591
22. Poplack SP, Tosteson TD, Kogel CA et al (2007) Digital breast tomosynthesis: Initial experience in 98 women with abnormal digital screening mammography. AJR Am J Roentgenol 189:616–623
23. Svahn T, Andersson I, Chakraborty D et al (2010) The diagnostic accuracy of dual-view digital mammography, single-view breast tomosynthesis, and a dual-view combination of breast tomosynthesis and digital mammography in a free-

response observer performance study. Radiat Prot Dosim 139:113–117
24. Teertstra HJ, Loo CE, van den Bosch MAAJ et al (2010) Breast tomosynthesis in clinical practice: initial results. Eur Radiol 20:16–24
25. Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M (2012) Two-view and single-view tomosynthesis versus full-field digital mammography: High-resolution X-ray imaging observer study. Radiology 262:788–796
26. Andersson I, Ikeda DM, Zackrisson S et al (2008) Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. Eur Radiol 18:2817–2825
27. Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Mattson S, Andersson I (2012) Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. Br J Radiol 85:e1074–e1082
28. Skaane P, Gullien R, Bjørndal H et al (2012) Digital breast tomosynthesis (DBT): initial experience in a clinical setting. Acta Radiol 53:524–529
29. Michell MJ, Iqbal A, Wasan RK et al (2012) A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis. Clin Radiol 67:976–981
30. Skaane P, Young K, Skjennald A (2003) Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading—Oslo I study. Radiology 229:877–884
31. Skaane P, Skjennald A (2004) 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 232:197–204
32. FDA Radiological Devices Panel Meeting, October 24, 2012, PMA application #P08003/S001http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/RadiologicalDevicesPanel/ucm299053.htm
33. Ruth C, Smith A, Stein J. US Patent 7760924. System and method for generating a 2D image from a tomosynthesis data set. Available at: http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=7760924.PN.&OS=PN/7760924&RS=PN/77609. Accessed 25 October 2011
34. Gur D, Zuley ML, Anello MI et al (2011) Dose reduction in digital breast tomosynthesis (DBT) screening using synthetically reconstructed projection images: An observer performance study. Acad Radiol 19:166–171
35. Gennaro G, Hendrick RE, Ruppel P et al (2012) Performance comparison of single-view digital breast tomosynthesis plus single-view digital mammography with two-view digital mammography. Eur Radiol 23:664-672
36. Gur D, Bandos AI, Rockette HE et al (2011) Localized detection and classification of abnormalities on FFDM and tomosynthesis examinations rated under an FROC paradigm. AJR Am J Roentgenol 196:737–741
37. Zuley ML, Bandos AI, Abrams GS et al (2010) Time to diagnosis and performance levels during repeat interpretations of digital breast tomosynthesis: Preliminary observations. Acad Radiol 17:450–455