

Digital Mammography versus Digital Mammography Plus Tomosynthesis in Breast Cancer Screening: The Oslo Tomosynthesis Screening Trial

Per Skaane, MD, PhD • Andriy I. Bandos, PhD • Loren T. Niklason, PhD • Sofie Sebuødegård, MSc • Bjørn H. Østerås, MSc • Randi Gullien, MSc • David Gur, ScD • Solveig Hofvind, PhD

From the Division of Radiology and Nuclear Medicine, Oslo University Hospital, University of Oslo, Breast Imaging Center, PO Box 4950, Nydalen, 0424 Oslo, Norway (P.S., R.G.); Departments of Biostatistics (A.I.B.) and Radiology (D.G.), University of Pittsburgh, Pittsburgh, Pa; Retired, former employee of Hologic (L.T.N.); Department of Screening, Cancer Registry of Norway, Oslo, Norway (S.S., S.H.); Department of Diagnostic Physics, Institute of Clinical Medicine, Oslo University Hospital, University of Oslo, Oslo, Norway (B.H.Ø.); and Department of Health Sciences, Oslo Metropolitan University, Oslo, Norway (S.H.). Received October 19, 2018; revision requested November 13; final revision received December 12; accepted January 2, 2019. Address correspondence to P.S. (e-mail: PERSKA@ous-hf.no).

Study supported by Hologic.

Conflicts of interest are listed at the end of this article.

See also the editorial by Lång in this issue.

Radiology 2019; 00:1–8 • <https://doi.org/10.1148/radiol.2019182394> • Content codes: **BR** **OI**

Background: Digital breast tomosynthesis (DBT) is replacing digital mammography (DM) in the clinical workflow. Currently, there are limited prospective studies comparing the diagnostic accuracy of both examinations and the role of synthetic mammography (SM) and computer-aided detection (CAD).

Purpose: To compare the accuracy of DM versus DM + DBT in population-based breast cancer screening.

Materials and Methods: This prospective study, performed from November 2010 to December 2012, included 24 301 women (mean age, 59.1 years \pm 5.7 [standard deviation]) with 281 cancers, of which 51 were interval cancers. Each examination was independently interpreted with four reading modes: DM, DM + CAD, DM + DBT, and SM + DBT. Sensitivity and specificity were compared for DM versus DM + DBT, DM versus DM + CAD, DM + DBT versus SM + DBT, and DM versus DM + DBT at double reading. Reader-adjusted performance characteristics of reading modes were evaluated on the basis of pre-arbitration (initial interpretation) scores. Statistical analysis was based on cluster bootstrap analysis using 10 000 random resamples.

Results: Sensitivity was 54.1% (152 of 281) for DM and 70.5% (198 of 281) for DM + DBT. Reader-adjusted difference was 12.6% (95% confidence interval [CI]: 5.2%, 19.7%; $P = .001$). Specificity was 94.2% (false-positive fraction [FPF], 5.8%; 1388 of 24 020) for DM and 95.0% (FPF, 5.0%; 1209/24 020) for DM + DBT, with a reader-adjusted difference in FPF of -1.2% (95% CI: -1.7% , -0.7% ; $P < .001$). Sensitivity was 69.0% (194 of 281) for SM + DBT and 70.5% (198 of 281) for DM + DBT, with a reader-adjusted difference of 1.0% (95% CI: -6.2% , 8.5%; $P = .77$). Specificity was 95.4% (FPF, 4.6%; 1111 of 24 020) for SM + DBT and 95.0% (FPF, 5.0%; 1209 of 24 020) for DM + DBT, with reader-adjusted 95% CIs for FPF of 4.7%, 5.4% and 5.0%, 5.7%, respectively, and a difference of -0.3% (95% CI: -0.8% , 0.2%; $P = .23$). Differences in sensitivity and specificity with the addition of CAD were small and not significant ($P > .2$).

Conclusion: Addition of digital breast tomosynthesis to digital mammography resulted in significant gains in sensitivity and specificity. Synthetic mammography in combination with digital breast tomosynthesis had similar sensitivity and specificity to digital mammography in combination with digital breast tomosynthesis.

© RSNA, 2019

Mammography screening reduces breast cancer mortality through early detection of small node-negative cancers (1,2). Digital mammography (DM) has two inherent limitations: low sensitivity in dense breasts because of a “masking effect” caused by overlying parenchyma and low specificity because summation of normal parenchyma can simulate a lesion. Results from retrospective studies (3–5) and prospective trials (6–8) have confirmed the potential of digital breast tomosynthesis (DBT) to address these limitations.

Several studies implementing DBT in screening used “combo mode” DM + DBT (3–7). However, use of this mode results in a doubling of radiation dose. Synthetic mammography (SM) images are a potential solution to this challenge and require no additional radiation dose.

The purpose of our prospective Oslo Tomosynthesis Screening Trial (OTST) was to compare diagnostic accuracy for independent reading of DM to DM + DBT, addition of computer-aided detection (CAD) to DM, and use of SM instead of DM in combination with DBT for breast cancer screening.

Materials and Methods

Hologic (Marlborough, Mass) sponsored this study by providing equipment and financial support for additional radiologist readings. Authors had full control of all data. The trial was approved by the regional ethical committee (clinical trial number NCT01248546). Written informed consent was required from all participants.

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

CAD = computer-aided detection, CI = confidence interval, DBT = digital breast tomosynthesis, DM = digital mammography, FPF = false-positive fraction, OTST = Oslo Tomosynthesis Screening Trial, SM = synthetic mammography, TPF = true-positive fraction

Summary

Sensitivity and specificity of screening mammography are significantly improved with the addition of tomosynthesis to digital mammography.

Key Points

- Implementing digital breast tomosynthesis and digital mammography into population-based breast cancer screening improved screening sensitivity (70.5% vs 54.1%, $P = .001$) and specificity (95.0% vs 94.2%, $P < .001$) compared with these values for digital mammography alone.
- The use of synthetic mammography images in combination with tomosynthesis showed no substantial differences in sensitivity (69.0% vs 70.5%) and specificity (95.4% vs 95%) compared with digital mammography plus tomosynthesis.

Four previous reports have been published about the OTST. A preplanned interim analysis after 1 year resulted in two reports comparing DM versus DM + DBT (7,9). A publication after 2 years of screening ($n = 24\,901$, a different number than in the current report because of differences in exclusions of women imaged twice) compared two versions of SM versus DM + DBT (10). Note that these interim results were published prior to follow-up. Women were followed for 2 years after the conclusion of screening to determine interval cancers, sensitivity, and specificity. A publication focused on interval cancers compared results of double reading of DM + DBT in the OTST ($n = 24\,301$) with those of two previous rounds of screening using DM alone (11). The current analysis reports final results of the OTST, including the sensitivity and specificity of all four arms and comparison of four imaging modes. These sensitivity and specificity results have not been previously reported.

Study Population

Between November 22, 2010, and December 19, 2012, 59 009 invitation letters were sent and 34 740 screening examinations were performed (participation rate, 58.9%). Women aged 50–69 years (mean age, 59.1 years \pm 5.7 [standard deviation], which is a similar range to ages in other Norwegian counties) (12) attended population-based screening through BreastScreen Norway, which invites women by personal letter to undergo two-view mammography biennially (13). On their arrival for the scheduled examination, women were asked to participate. The selection of women was based solely on the availability of radiographers and imaging systems. The ethics committee did not allow recording of the reason women gave for declining to participate. The few women who declined to undergo DBT and women who were not asked to participate because of limitations related to the availability of radiographers or equipment underwent DM imaging and were excluded. Women with pacemakers, women who were unable to stand, and women with breast implants were excluded.

A total of 1603 women attended twice, and all second examinations were excluded (Fig 1). A group of 8824 women who underwent DM only, six women with screening-detected malignancies (lymphomas and metastases), two women with palpable breast cancer and normal mammographic scores, and one woman with a screening-detected local recurrence were excluded. Three women scheduled for assessment did not return and were excluded. Hence, 24 301 women underwent DM + DBT and represent the study population (Fig 1).

Imaging Techniques

Screening examinations were performed with commercially available systems capable of DM and DBT imaging (Dimensions; Hologic). Two views (craniocaudal and mediolateral oblique) of each breast were obtained. DM and DBT were performed during a single breast compression per view. The average glandular radiation dose was 1.58 mGy \pm 0.61 for DM, 1.95 mGy \pm 0.58 for DBT, and 3.53 mGy \pm 0.84 mGy for DM + DBT (7).

Image Interpretation

Before the commencement of the trial, participating radiologists received individualized personal training of approximately 4 hours that involved reviewing at least 100 DBT examinations enriched with cancers.

Examinations in each of the following modes were independently interpreted by four different radiologists in a batch mode by using four dedicated workstations: DM (arm A), DM + CAD (arm B), DM + DBT (arm C), and SM + DBT (arm D) (Fig 1). Hanging protocols have been described previously (7). The CAD system was ImageChecker 9.3 (Hologic). DBT reconstruction methods were the same as those approved by the Food and Drug Administration (FDA) for use in the United States. Two versions of SM reconstruction were used: a prototype version in year 1 and a method approved by the FDA in year 2 (10).

Each radiologist rated his or her findings per breast by using a five-point ordinal rating scale for the probability of cancer, where a score of 1 indicated negative or definitely benign findings; a score of 2, probably benign findings; a score of 3, indeterminate findings; a score of 4, probably malignant findings; and a score of 5, malignant findings. Scores 2–5 were considered positive scores, and mammographic features were specified for each positive score.

All examinations that received at least one score of 2 or greater in at least one arm were discussed at a consensus meeting. A minimum of two radiologists participated in consensus-based arbitration meetings, with scores from each of the four arms available, during which a binary decision was made to either dismiss the findings or invite the participant for diagnostic work-up.

Outcome Measures and Study End Points

Radiologists attending the consensus/arbitration meeting had DM and DBT studies available for review. Results after the consensus meeting were influenced by the availability of DBT. To minimize changes resulting from the use of DBT at the

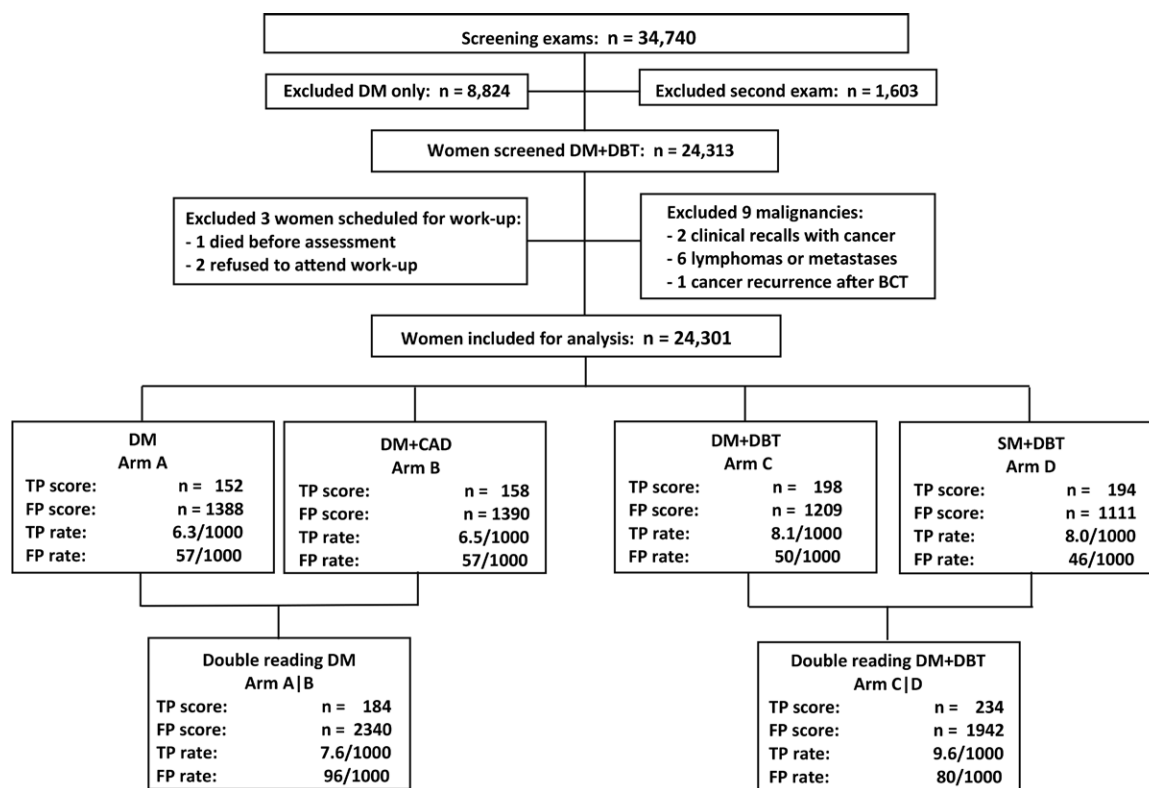


Figure 1: Flowchart of study participants and design, with true-positive (TP) and false-positive (FP) scores for each arm and double-reading modes. Arm A/B = arms A + B, Arm C/D = arms C + D, BCT = breast-conserving therapy, CAD = computer-aided detection, DBT = digital breast tomosynthesis, DM = digital mammography, SM = synthetic mammography.

consensus meeting, we focused our analysis on initial interpretations made before the consensus meeting.

The following four imaging modes were compared: (a) DM (arm A) versus DM + DBT (arm C), (b) DM (arm A) versus DM + CAD (arm B), (c) DM + DBT (arm C) versus SM + DBT (arm D), and (d) double reading of DM (arms A + B) versus double reading of DM + DBT (arms C + D).

Primary outcome measures were true-positive fraction (TPF) (sensitivity), false-positive fraction (FPF) ($1 - \text{specificity}$), and the predictive value of positive and negative scores. TPF was estimated as the fraction of women with verified cancer (where true-positive cases were verified with diagnostic work-up and false-negative cases were verified with 2-year follow-up). For simulated double-reading modes, a positive score in either of two combined modes for women with cancer in the same breast as the cancer was considered a true-positive finding. FPF was estimated as the fraction of women with a positive score among women who did not have cancer (verified by follow-up of all women without cancer). The predictive value of a positive score was estimated as the fraction of true-positive cases out of all cases with positive scores; the predictive value of a negative score was estimated as the fraction of non-cancer cases out of all cases without a positive score.

All cancers were confirmed through pathologic examination. The follow-up period was 24 months (from screening), and cases negative for cancer were verified to remain negative for this time period by querying the national cancer registry. An interval cancer was defined as a malignancy diagnosed clinically or at

imaging after a negative screening or assessment result but before the next scheduled screening examination. Tumor characteristics, molecular subtypes, and Ki-67 findings are given for invasive cancers, when available. Results obtained after the consensus meeting are reported.

Statistical Analyses

TPF (sensitivity) and FPF ($1 - \text{specificity}$) were estimated for each radiologist and each reading mode. Of nine participating radiologists (including P.S.), one reviewed only a small number of images in one reading mode (radiologist 9, arm B). A total of 56 cases reviewed by that radiologist were excluded from reader-adjusted analyses.

For each reading mode, we computed two estimates of sensitivity and specificity: pooled estimates based on all observations (eg, sensitivity as the number of cancer cases with positive scores in a given reading mode out of the total number of cancer cases) and reader-adjusted estimates represented by the average of radiologist-specific estimates. To alleviate the effect of imbalance in the numbers of cases evaluated by participating radiologists (which ranged from 564 to 5318 for individual study arms), statistical analysis was based on reader-adjusted estimates. Predictive values for positive and negative scores were estimated by using reader-adjusted estimates of sensitivity and specificity combined with the estimated cancer rate (1.16%; 281 of 24 301) in the entire cohort of women.

Primary assessment was focused on evaluating changes in reader-adjusted TPF and FPF with the addition of DBT (ie,

Table 1: Accuracy Characteristics of the Positive and Negative Scores and Their Changes across Different Imaging Modalities

Parameter	True-Positive Fraction: Sensitivity (%) [*]	False-Positive Fraction: 1 – Specificity (%)	Predictive Value of a Positive Score [*]	Predictive Value of a Negative Score
Reading modality				
Single reading of DM				
Arm A: DM	54.1 (152/281)	5.8 (1388/24 020)	9.9 (152/1540)	99.4 (22 632/22 761)
Reader-adjusted estimate [†]	56.8 (49.9, 63.3)	6.6 (6.2, 7.0)	9.2 (8.1, 10.3)	99.5 (99.4, 99.5)
Arm B: DM + CAD	56.2 (158/281)	5.8 (1390/24 020)	10.2 (158/1548)	99.5 (22 630/22 753)
Reader-adjusted estimate [†]	60.1 (54.1, 65.9)	6.8 (6.3, 7.2)	9.4 (8.4, 10.4)	99.5 (99.4, 99.6)
Double reading of DM				
Arms A + B	65.5 (184/281)	9.7 (2340/24 020)	7.3 (184/2524)	99.6 (21 680/21 777)
Reader-adjusted estimate [†]	65.8 (59.1, 72.2)	10.4 (9.9, 10.8)	6.9 (6.2, 7.6)	99.6 (99.5, 99.6)
Single reading of DM + DBT				
Arm C: DM + DBT	70.5 (198/281)	5.0 (1209/24 020)	14.1 (198/1407)	99.6 (22 811/22 894)
Reader-adjusted estimate [†]	69.4 (63.1, 75.1)	5.4 (5.0, 5.7)	13.1 (11.9, 14.4)	99.6 (99.5, 99.7)
Arm D: SM + DBT	69.0 (194/281)	4.6 (1111/24 020)	14.9 (194/1305)	99.6 (22 909/22 996)
Reader-adjusted estimate [†]	70.4 (64.3, 76.2)	5.1 (4.7, 5.4)	14.0 (12.6, 15.4)	99.6 (99.6, 99.7)
Double reading of DM + DBT				
Arms C + D	83.3 (234/281)	8.1 (1942/24 020)	10.8 (234/2176)	99.8 (22 078/22 125)
Reader-adjusted estimate [†]	83.6 (78.6, 87.5)	8.4 (8.0, 8.9)	10.4 (9.7, 11.1)	99.8 (99.7, 99.8)
Difference				
Arm C vs A (including DBT)	16.4 (46/281)	−0.7 (−179/24 020)	4.2	0.20
Reader adjusted [†]	12.6 (5.2, 19.7) [.0014]	−1.2 (−1.7, −0.7) [<.001]	4.0 (2.5, 5.4)	0.16 (0.07, 0.25)
Arm B vs A (including CAD)	2.1 (6/281)	0.0 (2/24 020)	0.3	0.03
Reader adjusted [†]	3.4 (−3.2, 10.1) [.33]	0.2 (−0.3, 0.7) [.47]	0.2 (−0.9, 1.4)	0.04 (−0.04, 0.12)
Arm D vs C (SM vs DM + DBT)	−1.4 (−4/281)	−0.4 (−98/24 020)	0.8	−0.02
Reader adjusted [†]	1.0 (−6.2, 8.5) [.77]	−0.3 (−0.8, 0.2) [.23]	0.8 (−0.8, 2.5)	0.01 (−0.08, 0.11)
Double reading of arms C + D vs arms A + B	17.8 (50/281)	−1.7 (−398/24 020)	3.5	0.23
Reader adjusted [†]	17.7 (11.8, 23.8) [<.001]	−1.9 (−2.5, −1.3) [<.001]	3.5 (2.7, 4.3)	0.23 (0.16, 0.31)
Additional cases				
Detected only with DBT (arm C vs A)	22.1 (62/281)	3.4 (821/24 020)	7.0 (62/883)	...
Detected only with CAD (arm B vs A)	11.4 (32/281)	4.0 (952/24 020)	3.3 (32/984)	...
Detected only with DBT (arms A + B vs arms C + D)	19.9 (56/281)	5.0 (1209/24 020)	4.4 (56/1265)	...

Note.—Unless otherwise specified, data in parentheses are raw data. Data in brackets are *P* values. CAD = computer-aided detection, DBT = digital breast tomosynthesis, DM = digital mammography, SM = synthetic mammography.

^{*} Based on location-corrected scores among all cancers.

[†] Data in parentheses are 95% confidence intervals for the reader-adjusted characteristic based on 10 000 bootstrap resamples.

between DM [arm A] and DM + DBT [arm C]). Secondary assessments focused on evaluating changes related to the use of CAD and SM instead of DM in combination with DBT. In addition, the effect of DBT addition in a double-reading setting was evaluated by comparing simulated double-reading of DM (arms A + B) with simulated double-reading of DM + DBT (arms C + D).

The significance of differences among reading modes was evaluated with 95% confidence intervals (CIs) and *P* values for testing the null hypothesis of equality by using a nonparametric cluster bootstrap approach with a woman as resampling unit (14). A total of 10 000 Monte Carlo bootstrap samples were used. Because of a previously conducted preplanned interim analysis, we used a significance threshold of .0264 for primary comparisons and a threshold of .05 for secondary assessments. Differences in the characteristics of detected cancers were tested by using the Fisher exact test (proc freq, SAS, version 9.4; SAS

Institute, Cary, NC). Differences in tumor size were tested by using the exact Wilcoxon test (proc npar1way, SAS, version 9.4).

Results

The mean age of the women imaged was 59.1 years \pm 5.7. There were 281 breast cancers, of which 51 were interval cancers. Table 1 shows estimates of TPF (sensitivity), FPF (1 – specificity), and the predictive value of positive and negative scores for each study arm and two simulated double-reading modes. Differences for four comparisons of interest of each outcome measure and reader-adjusted differences are shown.

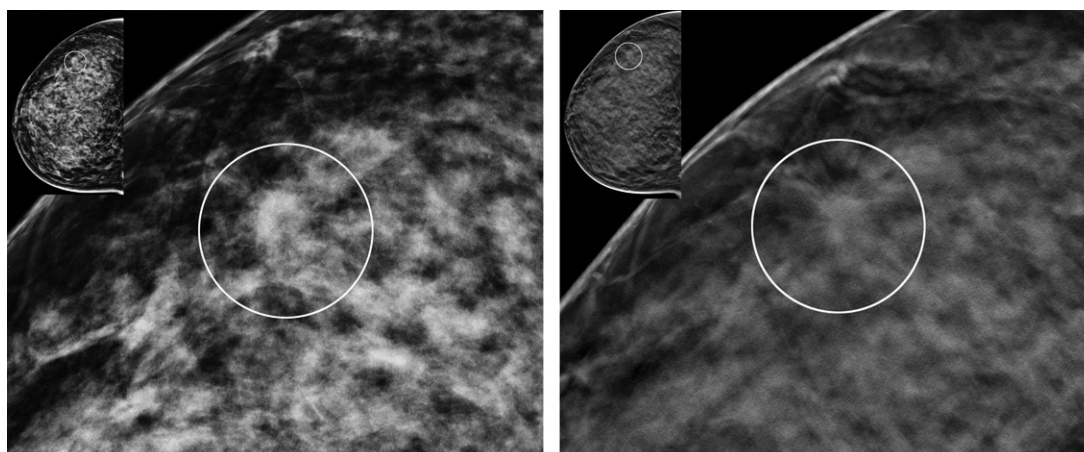
DM versus DM + DBT

Sensitivity was 54.1% (152 of 281) for DM and 70.5% (198 of 281) for DM + DBT, with reader-adjusted 95% CIs of 49.9%, 63.3% and 63.1%, 75.1%, respectively, and a difference of 12.6%

Table 2: Mammographic Features in False-Positive and True-Positive Cases

Mammographic Feature	DM (Arm A)		DM + CAD (Arm B)		DM + DBT (Arm C)		SM + DBT (Arm D)	
	FP	TP	FP	TP	FP	TP	FP	TP
Circumscribed mass	470 (33.9)	16 (10.5)	444 (31.9)	17 (10.8)	495 (40.9)	15 (7.6)	387 (34.8)	13 (6.7)
Spiculated mass	135 (9.7)	63 (41.4)	151 (10.9)	56 (35.4)	101 (8.4)	91 (46.0)	89 (8.0)	86 (44.3)
Architectural distortion	158 (11.4)	9 (5.9)	129 (9.3)	12 (7.6)	153 (12.7)	31 (15.7)	162 (14.6)	28 (14.4)
Asymmetric density	406 (29.3)	11 (7.2)	427 (30.7)	21 (13.3)	207 (17.1)	8 (4.0)	168 (15.1)	8 (4.1)
Calcifications	188 (13.5)	40 (26.3)	202 (14.5)	39 (24.7)	198 (16.4)	33 (16.7)	265 (23.9)	40 (20.6)
Calcifications and density	31 (2.2)	13 (8.6)	37 (2.7)	13 (8.2)	55 (4.5)	20 (10.1)	40 (3.6)	19 (9.8)
Total	1388	152	1390	158	1209	198	1111	194

Note.—Data are numbers of cases, with percentages in parentheses. CAD = computer-aided detection, DBT = digital breast tomosynthesis, DM = digital mammography, FP = false-positive, SM = synthetic mammography, TP = true-positive.



a.

b.

Figure 2: Screening mammography in 50-year-old asymptomatic woman with very dense breasts. **(a)** Digital mammogram of right breast in craniocaudal projection shows a small nonspecific round dense area (in circle) in the lateral part of the breast. Both readers concluded this was a normal finding (score of 1). **(b)** Tomosynthetic image in the same projection reveals a small spiculated mass (in circle) consistent with a small cancer (both reader scores were 4). Histologic examination revealed a grade 2 12-mm invasive ductal carcinoma (plus grade 3 ductal carcinoma in situ).

(95% CI: 5.2%, 19.7%; $P = .001$). The reader-adjusted difference was used for statistical analysis and differed slightly from the difference in pooled estimates. Specificity was 94.2% (FPE, 5.8%; 1388 of 24 020) for DM and 95.0% (FPE, 5.0%; 1209 of 24 020) for DM + DBT, with reader-adjusted 95% CIs for FPF of 6.2%, 7.0% and 5.0%, 5.7%, respectively, and a difference of -1.2% (95% CI: -1.7% , -0.7% ; $P < .001$). The predictive values of positive and negative scores also increased significantly with addition of DBT. The predictive value of a positive score for cases scoring positive only with DM + DBT was 7.0% (62 of 883).

DM versus DM + CAD

Sensitivity was 54.1% (152 of 281) for DM and 56.2% (158 of 281) for DM + CAD, with reader-adjusted 95% CIs of 49.9%, 63.3% and 54.1%, 65.9%, respectively, and a difference of 3.4% (95% CI: -3.2% , 10.1%; $P = .33$). Specificity was 94.2% (FPE, 5.8%; 1388 of 24 020) for DM and 94.2% (FPE, 5.8%; 1390 of 24 020) for DM + CAD, with reader-adjusted 95% CIs for FPF of 6.2%, 7.0% and 6.3%, 7.2%, respectively, and a difference of 0.2% (95% CI: -0.3% , 0.7%; $P = .47$).

DM + DBT versus SM + DBT

Sensitivity was 69.0% (194 of 281) for SM + DBT, compared with 70.5% (198 of 281) for DM + DBT, with reader-adjusted 95% CIs of 64.3%, 76.2% and 63.1%, 75.1%, respectively, and a difference of 1.0% (95% CI: -6.2% , 8.5%; $P = .77$). Specificity was 95.4% (FPE, 4.6%; 1111 of 24 020) for SM + DBT, compared with 95.0% (FPE, 5.0%; 1209 of 24 020) for DM + DBT, with reader-adjusted 95% CIs for FPF of 4.7%, 5.4% and 5.0%, 5.7%, respectively, and a difference of -0.3% (95% CI: -0.8% , 0.2%; $P = .23$).

DM versus DM + DBT with Double Reading

Sensitivity was 65.5% (184 of 281) for simulated double reading of DM and 83.3% (234 of 281) for simulated double reading of DM + DBT, with reader-adjusted 95% CIs of 59.1%, 72.2% and 78.6%, 87.5%, respectively, and a difference of 17.7% (95% CI: 11.8%, 23.8%, $P < .001$). Specificity was 90.3% (FPE, 9.7%; 2340 of 24 020) for DM and 91.9% (FPE, 8.1%; 1942 of 24 020) for DM + DBT, with reader-adjusted 95% CIs for FPF of 9.9%, 10.8% and 8.0%, 8.9%, respec-

tively, and a difference of -1.9% (95% CI: -2.5% , -1.3% ; $P < .001$).

Mammographic Features

Mammographic features in participants with positive scores are shown in Table 2. Additional true-positive cases identified with the addition of DBT were primarily spiculated masses and architectural distortions (Fig 2). For arms A and B (without DBT) there were more false-positive cases identified as asymmetric densities than in arms C and D. Arm D (with SM) had more false-positive calcification cases than arm C (with DM).

Histologic Findings and Molecular Subtype

Histopathologic findings in true-positive cases in simulated double reading for DM (arms A + B) versus DM + DBT (arms C + D) are presented in Table 3. There were 178 true-positive cases scored as positive with both DM and DM + DBT and 56 cases with positive scores only with DM + DBT. There were six cases scored as positive only with DM (results not shown). For each category, the number of cases with data available is listed. Invasive cancers with positive scores only with DM + DBT were lower grade ($P = .02$), with a higher fraction being lymph node negative ($P = .002$) and smaller in size ($P = .01$) compared with those detected with DM and DM + DBT. Cancers with positive scores only with DM + DBT were primarily molecular types luminal A and luminal B HER2 negative with lower Ki-67 values compared with those detected with DM (Table 4). A Ki-67 cut point of 14% was used in accordance with the literature (15,16).

Table 3: Histologic Findings in True-Positive Cases after Double Reading of DM Alone and DM + DBT

Histologic Finding	TP Cases in DM and DM + DBT (Arms A + B and C + D)	TP Cases in DM + DBT Only (Arms C + D)	<i>P</i> Value*
No. of cancers	178	56	
No. of cases of DCIS	33 (18.5)	4 (7.1)	
No. of invasive cancers	145 (81.5)	52 (92.9)	.06
Invasive cancers			
No. with histologic diagnosis	145	52	
IDC	79 (54.5)	25 (48.1)	.28
IDC and DCIS	45 (31.0)	15 (28.8)	
ILC	18 (12.4)	11 (21.2)	
No. of cancers assigned a grade	144	51	
Grade 1	43 (29.9)	24 (47.1)	
Grade 2	68 (47.2)	21 (41.2)	.02
Grade 3	33 (22.9)	6 (11.8)	
No. of cancers with known lymph node status	141	51	
Negative	110 (78.0)	49 (96.1)	
Positive	31 (22.0)	2 (3.9)	.002
No. of cancers with size measurements	135	51	
≤20 mm	109 (80.7)	49 (96.1)	
>20 mm	26 (19.3)	2 (3.9)	.01
Mean size (mm)	14.3	11.1	.01

Note.—Data are numbers of cases, with percentages in parentheses. DBT = digital breast tomosynthesis, DCIS = ductal carcinoma in situ, DM = digital mammography, IDC = invasive ductal carcinoma, ILC = invasive lobular carcinoma, TP = true-positive.

* Two-sided *P* value for Fisher exact test for comparing discrete distributions (proc freq, SAS, version 9.4; SAS Institute, Cary, NC); two-sided *P* value for the exact Wilcoxon test for comparing distributions of tumor size in millimeters (proc npar1way, SAS, version 9.4).

Table 4: Molecular Subtypes and Ki-67 Values for Invasive Cancers Detected by Using DM and DM + DBT Compared with Those Detected by Using Only DM + DBT

Parameter	TP Cases in DM and DM + DBT (Arms A + B and C + D)	TP Cases in DM + DBT Only (Arms C + D)
No. of invasive cancers	145	52
No. of cancers with molecular subtyping	135	49
Luminal A	62 (45.9)	31 (63.3)
Luminal B HER2 negative	59 (43.7)	16 (32.7)
Luminal B HER2 positive	3 (2.2)	2 (4.1)
HER2-positive enriched	2 (1.5)	0
Triple negative	9 (6.7)	0
No. of cancers with Ki-67 values available	117	42
Ki-67 value ≤ 14%	60 (51.3)	29 (69.0)
Ki-67 value 15%–25%	24 (20.5)	9 (21.4)
Ki-67 value > 25%	33 (28.2)	4 (9.5)
Mean Ki-67 value (%)	21.7	14.6

Note.—Data are numbers of cases, with percentages in parentheses. DBT = digital breast tomosynthesis, DM = digital mammography, TP = true-positive.

A radial scar was diagnosed in 28 women. Eleven women had a positive score (spiculated mass or architectural distortion) with DM (arms A + B), compared with 27 positive scores with DM + DBT (arms C + D). A total of 51 women were diagnosed with an interval cancer (11), and 10 had true-positive scores in at least one study arm.

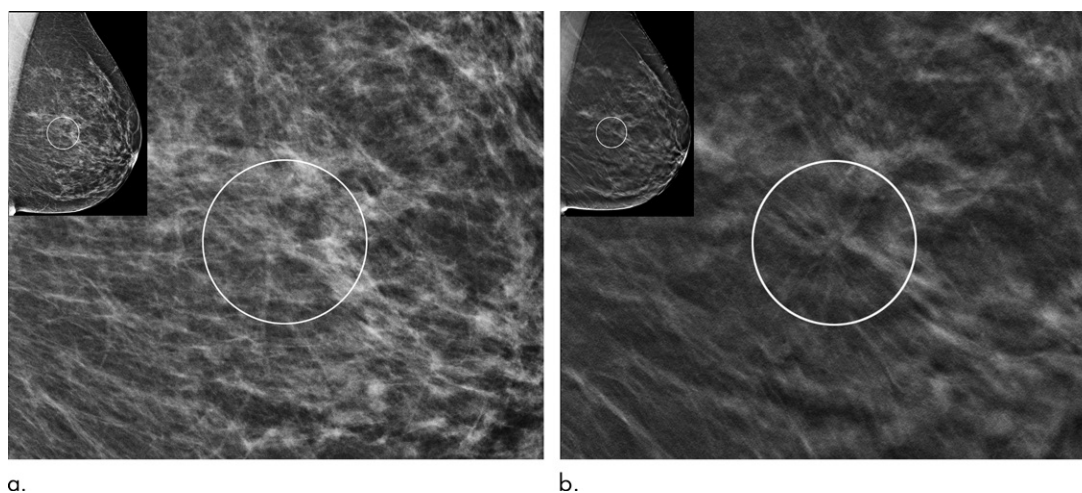


Figure 3: Screening mammography in 62-year-old asymptomatic woman. **(a)** Digital mammogram in mediolateral oblique view shows normal findings (in circle) (both independent digital mammography readers gave a negative score). **(b)** Tomosynthetic image in the same projection shows architectural distortion (in circle) (with both independent digital mammography plus tomosynthesis readers giving a positive score of 3). Histologic examination revealed a radial scar (with a small 4-mm papilloma) but no malignancy.

Post-Consensus Results

The numbers of screening-detected cancers after the consensus meeting for study arms A, B, C, and D were 147, 153, 194, and 189, respectively. A total of 938 women were recalled for assessment. The recall rate was 2.6% (623 of 24 301) for double reading of DM (arms A + B) and 3.4% (820 of 24 301) for double reading of DM + DBT (arms C + D). Double reading sensitivity after the consensus meeting was 62.6% (176 of 281) for DM and 80.8% (227 of 281) for DM + DBT. Specificity was 98.1% (FPE, 1.9%; 447 of 24 020) for DM and 97.5% (FPE, 2.5%; 593 of 24 020) for DM + DBT.

Discussion

We report the final results of the prospective population-based Oslo Tomosynthesis Screening Trial. Sensitivity improved significantly when DBT was added to the screening examination, from 54.1% (152 of 281) to 70.5% (198 of 281). Specificity also improved with DBT, from 94.2% (22 632 of 24 020) to 95.0% (22 811 of 24 020) for the single-reader mode. The predictive value of positive scores improved with the addition of DBT, from 9.9% (152 of 1540) to 14.1% (198 of 1407), while the predictive value of a negative score improved from 99.4% (22 632 of 22 761) to 99.6% (22 811 of 22 894). Similar gains were observed for simulated double-reading modes with the addition of DBT. Results for double reading are consistent with those of the prospective trial reported by Ciatto et al (6). Approximately two additional cancers per 1000 women screened were identified, while false-positives decreased with DBT. Although it was not an end point of our study, single reading with the addition of DBT resulted in higher sensitivity than double reading with DM alone.

Cases with positive scores were discussed at a consensus meeting in which DM, CAD, and DBT studies were available for each case. It was impractical to have two consensus meetings, one

without DBT for arms A and B and one with DBT for arms C and D. The availability of DBT distorted results primarily for the DM-only arms (A and B), especially for recall rate and specificity. We previously reported recall rates for two previous rounds of screening using DM, with an average recall rate of 4.2% (11). In this study, the post-arbitration recall rate for DM arms A + B was 2.6%. It is likely that the availability of DBT at consensus meeting allowed cases that would be recalled according to DM findings alone to be dismissed. To avoid distortions made after the consensus meeting, we focused on pre-consensus interpretations in this analysis.

Mammographic features of additional cancers identified with DM + DBT were predominantly spiculated masses or architectural distortions (Fig 2). Double reading of DM + DBT identified an additional 16 radial scars manifesting as architectural distortion, in accordance with the findings of a previous study (17), but this was accompanied by a larger gain in additional breast cancers classified as architectural distortion (Table 2, Fig 3). Analysis of false-positive scores revealed a reduction in asymmetric densities and suspicious spiculated masses with the addition of DBT (Table 2), unlike the increased recall of false-positive stellate distortions reported in a one-view DBT screening trial (18).

Cancers identified only with the addition of DBT were primarily low-grade, small, and node-negative cancers of molecular subtypes luminal A and luminal B HER2 negative with low Ki-67 values (19). A high fraction of cancers detected only with the addition of DBT were invasive lobular carcinomas (20%; 11 of 56). As in previous reports (20,21), the OTST did not show a reduction in the interval cancer rate (11). Further studies on interval cancer rates after DBT screening including additional rounds are required to determine the long-term benefits of DBT screening.

There have been conflicting reports on the value of CAD (22,23). Our study provides another data point. The estimated changes in both sensitivity and specificity were small and nonsignificant.

If SM could be used rather than DM with DBT, a screening examination could be performed at a radiation dose similar to that of DM. There have been positive reports on the use of SM in screening (12,24,25), but, to our knowledge, this study is the first large prospective trial comparing SM and DM in combination with DBT for the same set of cases. The estimated differences in sensitivity and specificity were negligible. Although levels of sensitivity with SM or DM combined with DBT were small, the numbers of cancers in this study were not large enough to exclude increases or decreases of 6%. Thus, additional studies are warranted.

Our study had limitations. First, it was a single-institution study involving equipment from a single vendor and a group of radiologists who were experienced in conventional mammography reading but had little experience using DBT before the study began. The scoring system used was different from that used in the United States, and the decision threshold for cases requiring further attention may have been different; however, the scoring system used provides an ordinal scale from negative to malignant, and the differences between modalities are highly likely to be similar. The common consensus meeting with availability of DBT was a limitation. Next, one arm in each main reading mode was modified (CAD in one DM arm and SM in one DBT arm), but we assume that the paired arms of the study for DM alone (arms A + B) and DM or SM + DBT (arms C + D) constitute sufficiently similar reading conditions to be considered double readings.

In conclusion, implementation of tomosynthesis in population-based breast cancer screening program significantly increased sensitivity and specificity. The use of synthetic mammography rather than digital mammography in combination with digital breast tomosynthesis resulted in little change in either sensitivity or specificity, which indicates that synthetic mammography might be a viable alternative to digital mammography when using digital breast tomosynthesis, although additional studies are needed.

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; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, P.S., L.T.N., S.H.; clinical studies, P.S., L.T.N., B.H.Ø., R.G., D.G.; statistical analysis, A.I.B., L.T.N., S.S., B.H.Ø., D.G.; and manuscript editing, P.S., A.I.B., L.T.N., B.H.Ø., R.G., D.G., S.H.

Disclosures of Conflicts of Interest: P.S. Activities related to the present article: received equipment and funding for additional case interpretations from Hologic. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. A.I.B. Activities related to the present article: the University of Pittsburgh was contracted to independently perform statistical analyses of the prospective tomosynthesis screening study. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. L.T.N. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a former employee of Hologic (retired in May 2016); has testified for Hologic in a tomosynthesis lawsuit between Hologic and Fuji Medical before the international trade commission; is the patent holder of the original breast tomosynthesis patent issued in 1997 to Massachusetts General Hospital and licensed to General Electric Medical (this patent has now expired); received royalties for patent on breast tomosynthesis before patent expiration; held stock and stock options in Hologic but divested after retirement in 2016. Other relationships: disclosed no relevant relationships. S.S. disclosed no relevant relationships. B.H.Ø. disclosed no relevant relationships. R.G. disclosed no relevant relationships. D.G. disclosed no relevant relationships. S.H. disclosed no relevant relationships.

References

1. Tabár L, Vitak B, Chen THH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology* 2011;260(3):658–663.
2. Njor S, Nyström L, Moss S, et al. Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies. *J Med Screen* 2012;19(Suppl 1):33–41.
3. Greenberg JS, Javitt MC, Katzen J, Michael S, Holland AE. Clinical performance metrics of 3D digital breast tomosynthesis compared with 2D digital mammography for breast cancer screening in community practice. *AJR Am J Roentgenol* 2014;203(3):687–693.
4. Conant EF, Beaber EF, Sprague BL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography compared to digital mammography alone: a cohort study within the PROSPR consortium. *Breast Cancer Res Treat* 2016;156(1):109–116.
5. Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA* 2014;311(24):2499–2507.
6. Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *Lancet Oncol* 2013;14(7):583–589.
7. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology* 2013;267(1):47–56.
8. Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. *Eur Radiol* 2016;26(1):184–190.
9. Skaane P, Bandos AI, Gullien R, et al. 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. *Eur Radiol* 2013;23(8):2061–2071.
10. Skaane P, Bandos AI, Eben EB, et al. Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images: comparison with digital breast tomosynthesis with full-field digital mammographic images. *Radiology* 2014;271(3):655–663.
11. Skaane P, Sebuodegård S, Bandos AI, et al. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. *Breast Cancer Res Treat* 2018;169(3):489–496.
12. Hofvind S, Hovda T, Holen AS, et al. Digital breast tomosynthesis and synthetic 2D mammography versus digital mammography: evaluation in a population-based screening program. *Radiology* 2018;287(3):787–794.
13. Hofvind S, Tsuruda K, Mangerud G, et al. The Norwegian Breast Cancer Screening Program, 1996–2016: celebrating 20 years of organized mammographic screening. In: *Cancer in Norway 2016 - cancer incidence, mortality, survival and prevalence in Norway*. Oslo, Norway: Cancer Registry of Norway, 2017.
14. Field CA, Welch AH. Bootstrapping clustered data. *J R Stat Soc Series B Stat Methodol* 2007;69(3):369–390.
15. Bustreo S, Osella-Abate S, Cassoni P, et al. Optimal Ki67 cut-off for luminal breast cancer prognostic evaluation: a large case series study with a long-term follow-up. *Breast Cancer Res Treat* 2016;157(2):363–371.
16. Healey MA, Hirko KA, Beck AH, et al. Assessment of Ki67 expression for breast cancer subtype classification and prognosis in the Nurses' Health Study. *Breast Cancer Res Treat* 2017;166(2):613–622.
17. Alshafei TI, Nguyen JV, Rochman CM, Nicholson BT, Patrie JT, Harvey JA. Outcome of architectural distortion detected only at breast tomosynthesis versus 2D mammography. *Radiology* 2018;288(1):38–46.
18. Lång K, Nergård M, Andersson I, Rosso A, Zackrisson S. False positives in breast cancer screening with one-view breast tomosynthesis: an analysis of findings leading to recall, work-up and biopsy rates in the Malmö Breast Tomosynthesis Screening Trial. *Eur Radiol* 2016;26(11):3899–3907.
19. Kim JY, Kang HJ, Shin JK, et al. Biologic profiles of invasive breast cancers detected only with digital breast tomosynthesis. *AJR Am J Roentgenol* 2017;209(6):1411–1418.
20. Bahl M, Gaffney S, McCarthy AM, Lowry KP, Dang PA, Lehman CD. Breast cancer characteristics associated with 2D digital mammography versus digital breast tomosynthesis for screening-detected and interval cancers. *Radiology* 2018;287(1):49–57.
21. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. *JAMA Oncol* 2016;2(6):737–743.
22. Gilbert FJ, Astley SM, Gillan MGC, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med* 2008;359(16):1675–1684.
23. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175(11):1828–1837.
24. Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. *Lancet Oncol* 2016;17(8):1105–1113.
25. Caumo F, Zorzi M, Brunelli S, et al. Digital breast tomosynthesis with synthesized two-dimensional images versus full-field digital mammography for population screening: outcomes from the Verona Screening Program. *Radiology* 2018;287(1):37–46.