

Assessing Radiologist Performance Using Combined Digital Mammography and Breast Tomosynthesis Compared with Digital Mammography Alone: Results of a Multicenter, Multireader Trial¹

Elizabeth A. Rafferty, MD
Jeong Mi Park, MD
Liane E. Philpotts, MD
Steven P. Poplack, MD
Jules H. Sumkin, MD
Elkan F. Halpern, PhD
Loren T. Niklason, PhD

¹From the Department of Radiology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114 (E.A.R., E.F.H.); Department of Radiology, University of Iowa Hospital, Iowa City, Iowa (J.M.P.); Department of Radiology, Yale University School of Medicine, Yale-New Haven Hospital, New Haven, Conn (L.E.P.); Department of Radiology, Dartmouth-Hitchcock Medical Center, Lebanon, NH (S.P.P.); Department of Radiology, Magee-Women's Hospital, Pittsburgh, Pa (J.H.S.); and Hologic, Bedford, Mass (L.T.N.). Received April 5, 2012; revision requested May 11; revision received June 19; accepted July 12; final version accepted July 26. **Address correspondence to E.A.R.** (e-mail: erafferty@partners.org).

© RSNA, 2012

Purpose:

To compare radiologists' diagnostic accuracy and recall rates for breast tomosynthesis combined with digital mammography versus digital mammography alone.

Materials and Methods:

Institutional review board approval was obtained at each accruing institution. Participating women gave written informed consent. Mediolateral oblique and craniocaudal digital mammographic and tomosynthesis images of both breasts were obtained from 1192 subjects. Two enriched reader studies were performed to compare digital mammography with tomosynthesis against digital mammography alone. Study 1 comprised 312 cases (48 cancer cases) with images read by 12 radiologists; study 2, 312 cases (51 cancer cases) with 15 radiologists. Study 1 readers recorded only that an abnormality requiring recall was present; study 2 readers had additional training and recorded both lesion type and location. Diagnostic accuracy was compared with receiver operating characteristic analysis. Recall rates of noncancer cases, sensitivity, specificity, and positive and negative predictive values determined by analyzing Breast Imaging Reporting and Data System scores were compared for the two methods.

Results:

Diagnostic accuracy for combined tomosynthesis and digital mammography was superior to that of digital mammography alone. Average difference in area under the curve in study 1 was 7.2% (95% confidence interval [CI]: 3.7%, 10.8%; $P < .001$) and in study 2 was 6.8% (95% CI: 4.1%, 9.5%; $P < .001$). All 27 radiologists increased diagnostic accuracy with addition of tomosynthesis. Recall rates for noncancer cases for all readers significantly decreased with addition of tomosynthesis (range, 6%–67%; $P < .001$ for 25 readers, $P < .03$ for all readers). Increased sensitivity was largest for invasive cancers: 15% and 22% in studies 1 and 2 versus 3% for in situ cancers in both studies.

Conclusion:

Addition of tomosynthesis to digital mammography offers the dual benefit of significantly increased diagnostic accuracy and significantly reduced recall rates for noncancer cases.

© RSNA, 2012

Supplemental material: <http://radiology.rsna.org/lookup/suppl/doi:10.1148/radiol.12120674/-/DC1>

Multiple randomized controlled trials have demonstrated that substantial reduction in breast cancer mortality can be realized through mammographic screening (1–5). With the implementation of digital mammography, additional diagnostic accuracy can be achieved for specific subgroups of women, presumably from its superior ability to depict cancers in dense breast tissue (6).

However despite its clearly documented benefit, it is well recognized that mammography is imperfect. As many as 20%–30% of breast cancers will not be detected on a mammogram (6,7). One of the factors negatively affecting the performance of mammography is breast density. Mammographic sensitivity decreases with increasing parenchymal density (6–9). On a two-dimensional mammographic projection, radiographically dense structures can be superimposed, potentially obscuring cancers. Conversely, these same overlapping structures can result in summation artifacts that mimic mammographic

abnormalities prompting false-positive recalls.

Breast tomosynthesis is a digital mammographic technique that permits individual planes of the breast to be visualized while reducing the impact from overlapping tissue (10). Unlike conventional digital mammography, in which each image is created from a single x-ray exposure, tomosynthesis images are reconstructed from a series of low-dose exposures as the x-ray source moves in an arc or linear trajectory above the breast. The resultant imaging data set minimizes the effect of overlapping structures, affording tomosynthesis the potential to enhance both the sensitivity and specificity of mammographic imaging.

Prior investigations of breast tomosynthesis have reported potential value of the technique in the diagnostic setting (11,12). Others have focused on feature visibility by using tomosynthesis imaging compared with conventional mammography (13,14). Two small retrospective observer studies showed recall rate reductions of 42% and 30% when using breast tomosynthesis combined with digital mammography compared with digital mammography alone (12,15). Of note, use of breast tomosynthesis alone did not result in a significant reduction in recall rate (15). Gennaro and colleagues (16) compared single-view tomosynthesis to two-view digital mammography in 200 women and found no significant difference in reader accuracy. However, to our knowledge, no multi-institutional trials comparing two-view tomosynthesis combined with digital mammography versus digital mammography alone have been reported. In this study, we compare radiologists' diagnostic accuracy and recall rates using breast tomosynthesis combined with

digital mammography in comparison with digital mammography alone.

Materials and Methods

One author (L.T.N.) is an employee of Hologic (Bedford, Mass), and one author (E.F.H.) is a statistical consultant for Hologic. Mammography review workstations for the reader study, as well as a grant for image collection and equipment, were provided to each of the five participating sites by Hologic. Authors without industry conflict of interest (E.A.R., J.M.P., L.E.P., S.P.P., and J.H.S.) had control of the data and written material for submission.

Study Design

The study protocol was approved by the institutional review boards of the five participating sites and was Health Insurance Portability and Accountability Act compliant. Women presenting for screening mammography or for breast biopsy were invited to participate and gave written informed consent. Participants underwent both digital mammography and tomosynthesis imaging of both breasts in the mediolateral oblique and craniocaudal positions. At each accruing site, a breast imager read the digital mammographic images and a second breast imager trained in tomosynthesis interpretation read the



Advances in Knowledge

- Two reader studies demonstrated a consistent and statistically significant gain in diagnostic accuracy (6.8% and 7.2% for the two studies) when breast tomosynthesis was added to conventional digital mammography.
- A significant reduction in recall rates for noncancer cases was demonstrated with the addition of breast tomosynthesis for all 12 radiologists participating in reader study 1 (mean reduction, 38.6%) and all 15 radiologists participating in reader study 2 (mean reduction, 17.1%).
- The addition of tomosynthesis resulted in large and significant improvement in area under the receiver operating characteristic curve for noncalcification cases (reader study 1, 8.8%; reader study 2, 10.4%), while for calcification cases the improvement was smaller and not significant.

Implication for Patient Care

- Tomosynthesis imaging may improve breast cancer detection while reducing recall rates at mammographic screening.

Published online before print

10.1148/radiol.12120674 **Content codes:**  

Radiology 2013; 266:104–113

Abbreviations:

AUC = area under the curve
 BI-RADS = Breast Imaging Reporting and Data System
 CI = confidence interval
 POM = probability of malignancy
 ROC = receiver operating characteristic

Author contributions:

Guarantor of integrity of entire study, E.A.R.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, E.A.R., S.P.P., L.T.N.; clinical studies, E.A.R., J.M.P., L.E.P., S.P.P., J.H.S., L.T.N.; statistical analysis, E.F.H., L.T.N.; and manuscript editing, all authors

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

tomosynthesis images for each participant presenting for screening; the respective breast imager was blinded to the results of the other modality. The investigational examination images of the participants presenting for biopsy were not prospectively interpreted.

Radiologists initially scored each study without access to prior imaging or clinical history and then scored the study, integrating this information. The initial score was used to classify the case as either negative (defined when both readers scored the case as a “nonrecall”) or recall (defined when one or both of the radiologists scored the case as a recall). Recall of screening patients for additional evaluation proceeded on the basis of the final recommendation of either reader. Diagnostic evaluation in recalled screening patients was undertaken in accordance with standard clinical practice. Lesion type and location were recorded for all actionable findings. After complete evaluation, each case was classified into one of four categories: malignant biopsy, benign biopsy, negative screening, and recalled screening. This categorization was used for the purpose of guiding case enrichment in subsequent reader studies. Subjects were monitored for up to an additional 2 years after enrollment. Two retrospective reader studies comparing digital mammography to digital mammography plus tomosynthesis were performed with the goal of comparing overall diagnostic accuracy and recall rates of noncancer cases for the two methods. Diagnostic accuracy was measured by using multireader receiver operating characteristic (ROC) analysis. Recall rates for noncancer screening cases were assessed for each reader.

Patient Population

A total of 1192 subjects were recruited from five sites between July 2006 and May 2007, of whom 997 subjects (780 screening cases and 217 biopsy cases) had complete imaging data sets and passed quality control review and were thus eligible for analysis (Fig 1).

Imaging Methods

Participants underwent digital mammography with a commercially available

system (Selenia; Hologic, Bedford, Mass). Tomosynthesis images were obtained by using an investigational tomosynthesis system (Hologic) utilizing a tungsten tube with 15° tube motion, 0.7 mm aluminum filtration, 11 projection images, a 10-second acquisition time, and a manual technique designed to match radiation dose to that delivered by the digital mammography system. Both imaging studies were acquired on the same day. With use of this protocol, the total dose for the combined digital mammography and tomosynthesis examinations was

two times that of digital mammography alone, although the combined dose was still less than the U.S. Food and Drug Administration’s limit for a single mammogram (17).

Determination of Reference Standard

Cases of women undergoing biopsy with malignant disease results were considered positive. Cases of women with concordant benign biopsy results and women not undergoing biopsy with no evidence of breast malignancy after 1 year of clinical follow-up were considered negative. Any case in which a

Figure 1

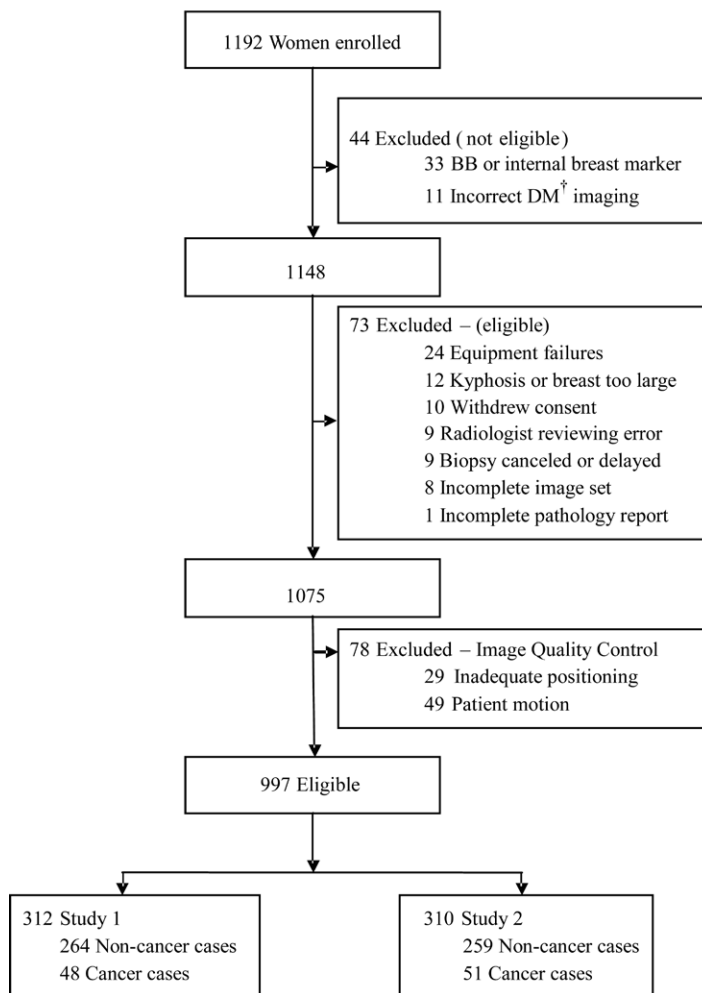


Figure 1: Subject imaging and case selection for breast tomosynthesis with digital mammography (DM). Women were excluded if they had a history of breast cancer, previous surgical biopsy, or an implanted tissue marker within the breast. Also excluded were women with breast implants and those who were pregnant or lactating.

woman was diagnosed with breast cancer within 365 days of enrollment was considered positive regardless of initial classification.

Reader Study Case Selection

To assess the two primary endpoints of diagnostic accuracy and noncancer recall rate, a subset of the eligible cases was selected for each of two reader studies. The number of cases included from each of the four case types was chosen based on a power analysis designed to provide 80% power to detect a 0.05 difference in the mean area under the curve (AUC) and a 20% difference in recall rates for individual readers. The resulting number of cases for each case type in each reader study is shown (Table 1).

All eligible cases with documented malignancy were included in both of the reader studies (48 cases in reader study 1 and 51 in reader study 2); five were accrued from the screening population, and the remainder came from the biopsy group. A breakdown of the cancer cases is shown (Table 2). Noncancer cases (benign biopsy, recalled screening, and negative screening) were selected randomly from the eligible cases in each category. A separate, independent randomization of noncancer cases was used for each reader study.

Reader Studies

Fourteen radiologists who currently interpret mammograms in clinical practice were invited to participate in the first reader study, and 15 different radiologists were invited to participate in the second reader study. None of the radiologists had prior experience in the interpretation of tomosynthesis images. Training in the interpretation of breast tomosynthesis images was provided by an experienced reader. The training for reader studies 1 and 2 consisted of review of approximately 150 cases illustrating the appearance of normal tissue patterns, summation artifact, and benign and malignant lesions. When an analysis of reader study 1 suggested that readers had

inappropriately dismissed some circumscribed, lobulated masses as benign findings, training for reader study 2 was supplemented with three additional examples of circumscribed, lobulated lesions. These examples were not differentiated from the other training cases, but were simply mixed in with the other examples. A summary of the training was provided in written format. No cases utilized in the training sets were included in the respective reader studies. The readers were required to pass cancer detection and maximum recall rate thresholds for tomosynthesis imaging. Twelve of 14 readers completed the training and met inclusion thresholds in the first study. All 15 readers met criteria for inclusion in the second reader study.

Readers first scored the digital mammogram. They were then provided with the tomosynthesis images and scored the combined study. Three scores were recorded for each of the imaging methods. An initial Breast Imaging Reporting and Data System (BI-RADS) (18) score of 0 (recall), 1 (negative), or 2 (benign) was used to determine the recall rate. For studies scored as a recall (BI-RADS 0), the reader then provided a forced BI-RADS score of 1, 2, 3, 4, or 5 to indicate the most likely outcome based on the appearance of the finding. These scores were used to calculate diagnostic sensitivity, specificity, and positive and negative predictive values. A probability of malignancy (POM) score ranging from 0%–100% was also recorded for each case. The forced BI-RADS and POM scores were used to calculate ROC curves.

In addition to data recorded for reader study 1, radiologists in reader study 2 identified the involved breast and lesion type (calcification or noncalcification) for all actionable findings. These identifiers were used to confirm that readers had correctly identified a cancer.

Statistical Considerations

The study was prospectively designed to test the null hypotheses of equality of cancer detection rates (as determined

Table 1

Number of Cases for Each Case Type Used in Reader Studies 1 and 2

| Case Type | Eligible Cases | Reader | |
|-----------|----------------|---------|---------|
| | | Study 1 | Study 2 |
| Cancer | 51 | 48* | 51 |
| Recall | 248 | 141 | 138 |
| Benign | 166 | 48 | 47 |
| Negative | 532 | 75 | 74 |
| Total | 997 | 312 | 310 |

Note.—The mean age of the subjects was 51.7 years (range, 25–80 years) in study 1 and 53.5 years (range, 25–87 years) in study 2. One-year follow-up was available for all cases except 9.6% and 9.7% of cases lost to follow-up in studies 1 and 2, respectively. BI-RADS breast density scores of 1, 2, 3, or 4 were respectively assigned for 22, 149, 134, and seven cases in study 1 and 24, 135, 146, and five cases in study 2.

* Forty-eight of 51 total cancer cases had complete pathology reports at the time of reader study 1.

by POM score in a multireader, multi-case ROC analysis) and equality of noncancer recall rates for the individual readers. Individual recall rates for cancers and noncancers were calculated for each imaging method by calculating the fraction of cases scored as BI-RADS 0 for each radiologist. All noncancer cases were used in the recall analysis in both studies. The McNemar test was used for comparison of individual radiologist's recall rates.

Diagnostic sensitivity, specificity, and positive and negative predictive values were compared by using a BI-RADS score of 4 or 5 considered as positive and a BI-RADS score of 1, 2, or 3 considered as negative. The McNemar test was used for comparison of individual reader's scores.

Post hoc comparison of pooled recall rates, diagnostic sensitivity, and diagnostic specificity across all readers was performed by using logistic regression analysis with terms for the mode (digital mammography vs digital mammography plus tomosynthesis), reader, case, and reader by mode interaction.

ROC analysis was performed by using DBM MRMC 2.2 software (19,20). AUC was compared for each imaging method. Two-sided *P* values

Table 2

Characteristics of Cancer Cases

| Reader Study No. | DCIS (Noninvasive)* | Median Size (mm) [†] | IDC ± DCIS | IDC and ILC | ILC | Papillary | Adenoid Cystic | Total Invasive Cancers* | Median Size (mm) [†] | Total Cancers [‡] |
|------------------|---------------------|-------------------------------|------------|-------------|-----|-----------|----------------|-------------------------|-------------------------------|----------------------------|
| 1 | 16 (33) | 9.8 (3–36) | 24 | 4 | 1 | 2 | 1 | 32 (67) | 14.0 (6–34) | 48 |
| 2 | 16 (31) | 9.8 (3–36) | 25 | 5 | 2 | 2 | 1 | 35 (69) | 13.0 (6–34) | 51 |

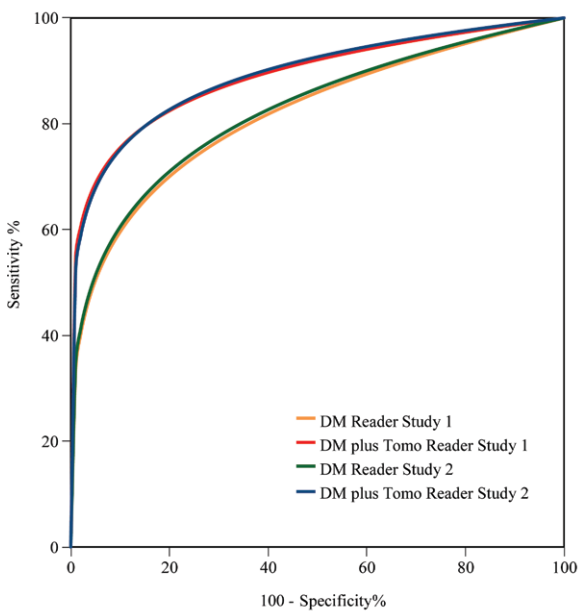
Note.—The mean age of the subjects with cancer was 56.8 years (range, 29–80 years) in study 1 and 56.9 years (range, 29–80 years) in study 2. DCIS = ductal carcinoma in situ, IDC = invasive ductal carcinoma, IDC ± DCIS = all IDCs, some of which may also have a component of DCIS, ILC = invasive lobular carcinoma.

* Numbers in parentheses are percentages.

[†] Size of invasive cancers and ductal carcinoma in situ were measured from digital mammograms, except for three cases in which size could not be determined; for these, the sonography (two cases) or magnetic resonance imaging (one case) report was used to determine tumor size. Numbers in parentheses are the range.

[‡] For reader study 1, 48 cancer cases had complete data; five of the cancer cases were from the screening group (one detected at tomosynthesis only, one at digital mammography only, and three detected at both examinations), and the remainder were from the biopsy group. For reader study 2, pathology reports were available for an additional three cancer cases from the biopsy group, for a total of 51 cancer cases.

Figure 2



Area under the ROC curve

| | DM [†] plus Tomo | | Difference | p-value | 95% CI |
|----------------|---------------------------|------|------------|---------|-----------|
| | DM | Tomo | | | |
| Reader Study 1 | 82.1 | 89.4 | 7.2 | <0.001 | 3.7, 10.8 |
| Reader Study 2 | 82.8 | 89.5 | 6.8 | <0.001 | 4.1, 9.5 |

were reported; $P < .05$ was considered to indicate a significant difference. ROC curves presented used the binormal model. Analysis assumed random readers and random cases.

Owing to missing reader scores, 308 and 303 cases could be used for the

ROC analysis for reader studies 1 and 2, respectively. All of the cancer cases had complete scores in both studies.

Logistic regression and McNemar test were performed by using statistical software (SAS, version 9.1; SAS Institute, Cary, NC).

Results

Cancer Cases

Of the malignant cases, 16 cases were ductal carcinoma in situ alone; the remainder were invasive or combined invasive and in situ cancers (Table 2). For invasive cancers, the median size was 14 and 13 mm for reader studies 1 ($n = 32$) and 2 ($n = 35$), respectively.

Diagnostic Accuracy

In both studies, digital mammography plus tomosynthesis demonstrated superior diagnostic accuracy compared with digital mammography alone, as shown by significant difference in the AUC in reader study 1 (AUCΔ = 7.2%; 95% confidence interval [CI]: 3.7%, 10.8%; $P < .001$) and reader study 2 (AUCΔ = 6.8%; 95% CI: 4.1%, 9.5%; $P < .001$) (Fig 2). For each of the 27 readers in reader studies 1 and 2, the AUC for the combined modality was greater than the AUC for digital mammography alone. At 1 year follow-up, no interval cancers had been reported in any of the reader study subjects.

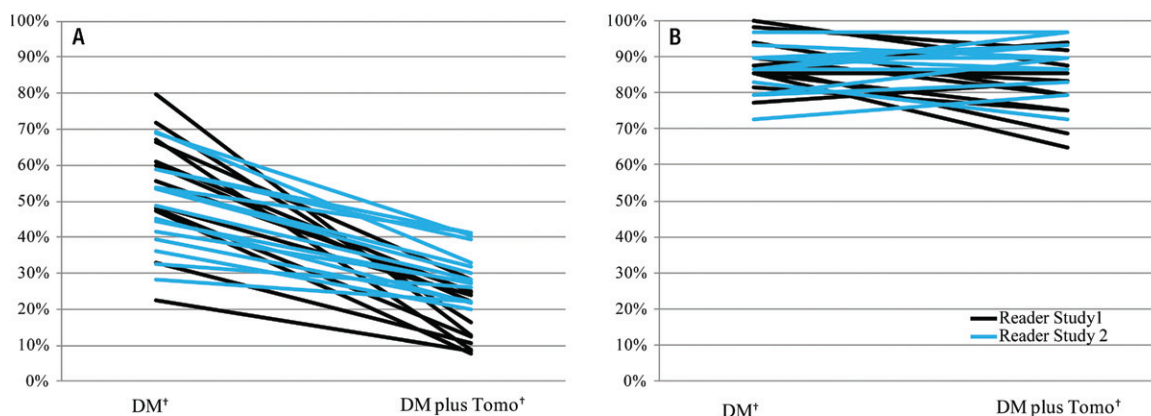
Diagnostic sensitivity and positive and negative predictive values increased with addition of tomosynthesis in reader study 1 by 11%, 13%, and 2% and in reader study 2 by 16%, 3%, and 3%, respectively (Table 3). Diagnostic specificity increased by 5% in reader study 1 and decreased by 2% in reader study 2. The increase in sensitivity was largest for invasive cancers, with

Table 3
Diagnostic Sensitivity, Specificity, and Positive and Negative Predictive Values

| Parameter | Reader Study 1 | | | Reader Study 2 | | |
|--------------------------------------|---------------------|--------------------------------|------------|---------------------|--------------------------------|------------|
| | Digital Mammography | Mammography plus Tomosynthesis | Difference | Digital Mammography | Mammography plus Tomosynthesis | Difference |
| Sensitivity (%) | 65.5 | 76.2 | 10.7 | 62.7 | 78.7 | 16.0 |
| Specificity (%) | 84.1 | 89.2 | 5.1 | 86.2 | 84.5 | -1.7 |
| Positive predictive value (%) | 42.9 | 56.2 | 13.3 | 47.3 | 50.1 | 2.8 |
| Negative predictive value (%) | 93.0 | 95.4 | 2.4 | 92.1 | 95.3 | 3.2 |
| Sensitivity for invasive cancers (%) | 63.8 | 78.6 | 14.8 | 60.6 | 82.3 | 21.7 |
| Sensitivity for in situ cancers (%) | 68.8 | 71.4 | 2.6 | 67.5 | 70.8 | 3.3 |

Note.—Cases with BI-RADS scores of 4 and 5 were considered positive and cases with BI-RADS scores of 1, 2, and 3 were considered negative.

Figure 3



Recall Rates (Average of Readers)

| Case Type | Reader Study | DM | | | DM plus Tomo | | |
|------------|--------------|-------|---------------|-------|--------------|---------------|------|
| | | Mean | Range | SD | Mean | Range | SD |
| Non-Cancer | 1 | 55.1% | 22.3% - 79.8% | 16.3% | 16.7% | 7.6% - 28.4% | 7.6% |
| | 2 | 48.8% | 28.2% - 69.1% | 12.3% | 30.1% | 19.8% - 41.3% | 7.6% |
| Cancer | 1 | 87.2% | 77.0% - 100% | 6.5% | 80.4% | 64.6% - 93.8% | 8.8% |
| | 2 | 84.8% | 76.0% - 92.2% | 6.1% | 85.7% | 78.0% - 92.2% | 6.4% |

Figure 3: Recall rates for *A*, noncancer and, *B*, cancer cases for individual readers. *DM* = digital mammography, *SD* = standard deviation, *Tomo* = tomosynthesis.

increases of 15% and 22% in reader study 1 and 2, respectively, while the increases for in situ cancer were 3% in both studies.

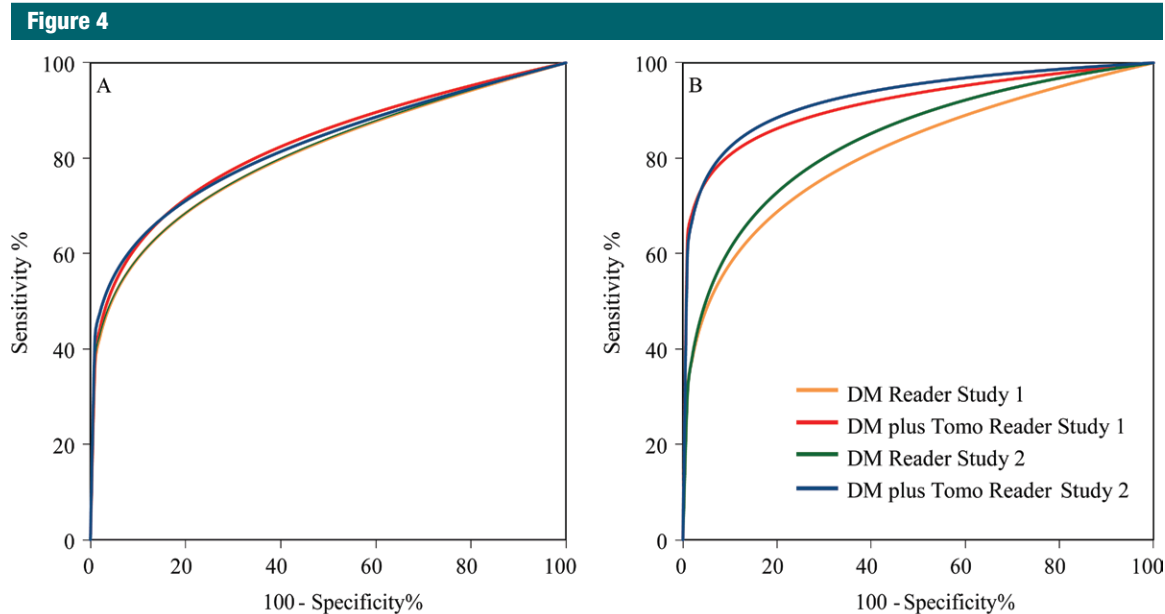
Overall diagnostic specificity improved for 12 of the 27 radiologists (statistically significant for eight), was equal for one, and decreased for 14 (statistically significant for six). Diagnostic sensitivity increased for 26 of 27

radiologists, and this difference was statistically significant for 10 radiologists. No radiologist showed a significant decrease in diagnostic sensitivity. The pooled logistic regression analysis did not demonstrate a significant change in diagnostic sensitivity (reader study 1, $P = .301$; reader study 2, $P = .545$) or specificity (reader study 1, $P = .546$; reader study 2, $P = .565$). Individual

results are shown in Tables E1 and E2 (online).

Recall Rate

Figure 3 presents individual recall rates for noncancer cases for both studies. Post hoc logistic regression analysis demonstrated a significant difference in recall rates of noncancers between the two methods ($P < .001$) but also a



Area under the ROC curve

| Case Type | Reader Study | DM plus [†] | | Difference | p-value | 95% CI |
|---------------------|--------------|----------------------|------|------------|---------|-----------|
| | | DM [†] | Tomo | | | |
| Calcification | 1 | 80.4 | 84.0 | 3.5 | 0.073 | -0.4, 7.4 |
| | 2 | 81.7 | 83.1 | 1.4 | 0.082 | -0.2, 2.9 |
| Non - Calcification | 1 | 80.7 | 91.2 | 10.4 | <0.001 | 4.7, 16.1 |
| | 2 | 84.2 | 93.0 | 8.8 | <0.001 | 5.1, 12.5 |

Figure 4: Pooled ROC curves for reader studies 1 and 2: A, calcification and, B, noncalcification cases. ROC curves were calculated from probability of malignancy scores except for calcification curves for study 1; forced BI-RADS scores were used for these curves because data were available for all 12 readers. The ROC program failed to produce an ROC curve using probability of malignancy scores for this subanalysis. For all other analyses, BI-RADS and POM ROC curves provided nearly equal results. There were 83 and 79 cases classified as calcification in studies 1 and 2, respectively; the remainder were classified as noncalcification. *DM* = digital tomography, *Tomo* = tomosynthesis.

significant ($P < .001$) interaction of the method with the reader. The prospectively defined analysis of the difference between the two modalities showed a significant reduction for every individual reader ranging from 6% to 67% ($P < .001$ for 25 of 27 readers, and $P < .03$ for all readers).

The difference in recall rates for cancers between the two modalities was not significant with use of post hoc logistic regression analysis ($P > .90$). In reader study 1 the interaction of method with reader was significant ($P < .001$), while in reader study 2 no significant interaction of method with reader was found ($P > .80$). In reader study 1, with use

of combined digital mammography and tomosynthesis, three readers exhibited nonsignificant increases in their cancer recall rates, one reader had the same recall rates, and eight readers showed decreases in their cancer recall rates, which for four readers were significant decreases. In reader study 2, no individual reader demonstrated a significant difference in cancer recall rates. In reader study 2, 3.1% of the cancer cases recalled on the basis of digital mammography were not correctly localized by the reader, while only 1.9% of cancer cases were incorrectly localized by using digital mammography combined with tomosynthesis.

Calcification versus Noncalcification

Mammographic abnormalities were classified as calcifications or noncalcifications by the interpreting radiologists at the accruing sites. When diagnostic accuracy is evaluated by calcification versus noncalcification imaging features, nearly all of the gain in reader performance is attributable to noncalcification cases (Fig 4). In both studies, there was a nonsignificant increase in diagnostic accuracy for calcification cases by using digital mammography plus tomosynthesis. The gain in diagnostic accuracy for noncalcification cases in both studies was, however, significant ($P < .001$). An example demonstrating the improved visibility of a noncalcification lesion is shown in Figure 5.

Figure 5

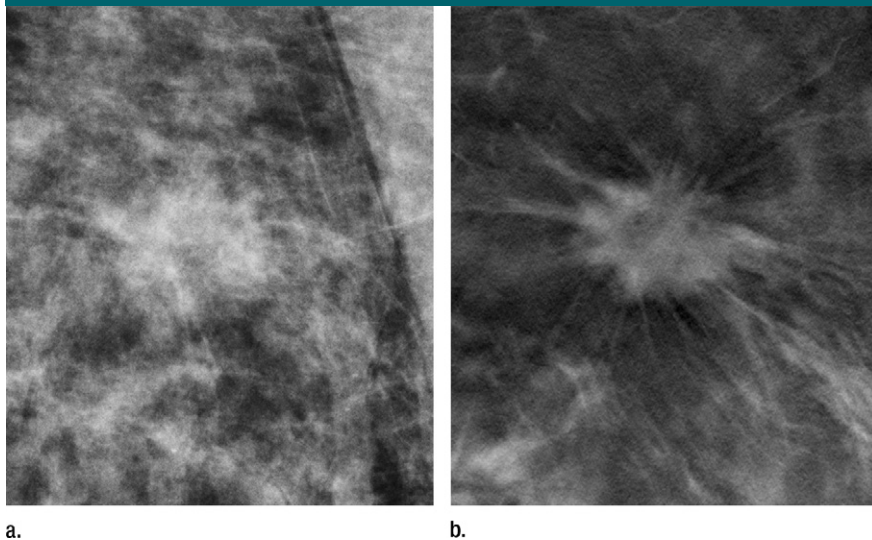


Figure 5: (a) Digital mammography and (b) tomosynthesis images of an invasive ductal carcinoma.

Discussion

Combining digital mammography with tomosynthesis offers interpretive advantages. The presence of the standard mammogram facilitates comparison with prior examinations and provides a comprehensive view of distributional features (particularly for calcifications) while the addition of tomosynthesis minimizes the effect of tissue overlap and allows better visualization of noncalcification features. Thus the relative strengths of the two modalities are retained with the combined approach; however, the addition of tomosynthesis to the standard mammogram represents additional radiation exposure to the patient. Investigational efforts are underway to replace the standard mammogram with a mammogram synthesized from the tomosynthesis images to reduce the dose.

With use of ROC analysis in two separate reader studies, radiologists significantly improved their diagnostic accuracy by reading digital mammography in conjunction with tomosynthesis compared with digital mammography alone. Across both studies, all 27 readers demonstrated improvement in their diagnostic accuracy. The

7.2% and 6.8% average gains in AUC for the two separate reader studies were consistent and significant. These gains are similar to the 7% gain reported by Gur and colleagues (15,21) in a smaller enriched reader study evaluating 125 cases.

In the screening setting, the decision to recall or not recall determines screening sensitivity and specificity. We also report diagnostic sensitivity and specificity based on the BI-RADS scale. This scale is typically used in the diagnostic setting and is a measure of the ability of radiologists to accurately predict the benign or malignant nature of lesions. Based on the BI-RADS classification, average diagnostic sensitivity increased with the addition of tomosynthesis in the two reader studies by 10.7% and 16.0%. In both reader studies, however, the screening sensitivity did not demonstrate significant improvement, presumably because the readers sometimes initially recalled cancer cases yet subsequently assigned them a low BI-RADS score (indicating that based on the available information, the finding warranted further evaluation but would likely ultimately prove to be benign or negative).

Almost all of the gains in diagnostic sensitivity realized with the combined modality were attributable to

the improved detection and characterization of invasive cancers. Clinically, a lesion presenting as calcifications is much more likely to represent noninvasive rather than invasive malignancy. Because tomosynthesis reduces tissue superimposition in the breast, its impact in rendering lesions more visible is most evident for masses, asymmetries, and areas of architectural distortion. The visibility of calcifications is degraded to a much lesser degree by overlapping tissue, thus one would not expect significant detection gains for lesions presenting as calcifications (12,14).

Furthermore, the use of the combined modalities in both reader studies produced a significant reduction in the recall rate of women who did not have cancer. In fact, all 27 readers in the two reader studies showed a significant reduction in noncancer recall rate. Clinically, such a reduction in recall rate can be expected to translate to a substantial number of unnecessary diagnostic tests being avoided. The number of false-positive studies resulting from screening mammography has come under recent scrutiny as one of the “harms” of mammography (22,23). Considering the relative emotional, financial, and clinical costs of screening versus diagnostic testing, the potential for tomosynthesis to reduce false-positive findings should provide multifaceted benefit for society.

While reduction in noncancer recalls represents a clear advantage, care must be taken to avoid misclassification of malignant lesions when tomosynthesis is added to digital mammography in interpretation. On review of the results of reader study 1, it was apparent that cancers manifesting as certain finding types, in particular circumscribed lobulated masses, were being inappropriately dismissed by some readers. In mammographic interpretation, radiologists often associate circumscribed masses with a benign or probably benign process. In tomosynthesis imaging however, circumscribed margins, particularly when associated with lobulated lesions, may be an indication of malignancy. This is important to emphasize

as clinical radiologists transition to interpreting tomosynthesis studies. Thus, although the training for reader studies 1 and 2 was identical in content, training for reader study 2 reinforced these principles in written format and with additional examples. Through closer adherence to prescribed training, nearly all of the readers in reader study 2 correctly classified circumscribed, lobulated lesions. The collective results of reader studies 1 and 2 emphasize the importance of effective training in the clinical environment to avoid inappropriate dismissal of some cancers.

Our study had several limitations. First, the reader studies were enriched, and results may be different from those of a true screening population. Further studies done in the clinical environment will be critical to confirm the performance of the modality. There was an inherent inclusion bias against tomosynthesis with respect to cancer detection in a screening population. Nearly all of the cancers were acquired in patients scheduled for biopsy and had been detected on conventional mammograms as part of standard screening evaluation. This biases the studies against tomosynthesis imaging because those cancers that could only be detected by tomosynthesis were not included; hence it is likely that our studies underestimate the potential gains in sensitivity that might occur in clinical practice.

Another limitation of the study was that prior imaging was not available to the readers to assist them in making their assessments. Additionally, the study was highly enriched with biopsy lesions, both malignant and nonmalignant, as well as with cases that had been identified for recall by the accruing sites. As a result of both these factors, the tendency to recall among the readers was likely higher than normally would be observed, and it is therefore also possible that the magnitude of reduction in recall rate realized will not be as substantial in clinical practice.

In conclusion, the addition of tomosynthesis to digital mammography offers the dual benefit of improved diagnostic accuracy and significant reduction in

false-positive recall rate thereby avoiding unnecessary additional testing and decreasing attendant anxiety, inconvenience, and cost for women.

Acknowledgments: We thank Hendrik J. Teertstra, MD, of the Netherlands Cancer Institute in Amsterdam, for providing cases used in the training sessions. We also thank Jennifer Bartoshevich, Kathleen Willison, and Lynne Jameson-Meehan for technical assistance.

Disclosures of Conflicts of Interest: E.A.R. No relevant conflict of interest to disclose. J.M.P. No relevant conflict of interest to disclose. L.E.P. No relevant conflict of interest to disclose. S.P.P. No relevant conflict of interest to disclose. J.H.S. Financial activities related to the present article: grants to institution from NIH and Komen. Financial activities not related to the present article: consultancy for Guidepoint Global, Morgan Stanley, MEDACorp; expert testimony in malpractice cases, Steven J. Forrey, Stege & Michelson Co., L.P.A., Moscarino & Treu LLP; payment for lectures including service on speakers bureaus, Educational Symposia (review course), Postgraduate Institute for Medicine (webinar), Innova Health System (breast seminar), and University of Florida (visiting professor); travel/accommodations/meeting expenses, Suros (scientific advisory board). Other relationships: none to disclose. E.F.H. Financial activities related to the present article: consultant and statistician for Hologic. Financial activities not related to the present article: expert testimony for Ameritox. Other relationships: none to disclose. L.T.N. Financial activities related to the present article: employee of Hologic. Financial activities not related to the present article: patent on breast tomosyntheses issued by Massachusetts General Hospital; stock in Hologic. Other relationships: none to disclose.

References

- Shapiro S, Venet W, Strax P, Venet L, Roesser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982;69(2):349-355.
- Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1985;1(8433):829-832.
- Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001;91(9):1724-1731.
- Tabár L, Vitak B, Chen HH, et al. The Swedish Two-County Trial twenty years later: updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am* 2000;38(4):625-651.
- Tabár L, Vitak B, Chen TH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology* 2011;260(3):658-663.
- 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(17):1773-1783.
- Holland R, Mravunac M, Hendriks JH, Bekker BV. So-called interval cancers of the breast: pathologic and radiologic analysis of sixty-four cases. *Cancer* 1982;49(12):2527-2533.
- Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* 1992;184(3):613-617.
- Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med* 2003;138(3):168-175.
- Niklason LT, Christian BT, Niklason LE, et al. Digital tomosynthesis in breast imaging. *Radiology* 1997;205(2):399-406.
- Teertstra HJ, Loo CE, van den Bosch MA, et al. Breast tomosynthesis in clinical practice: initial results. *Eur Radiol* 2010;20(1):16-24.
- Poplack SP, Tosteson TD, Kogel CA, Nagy HM. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. *AJR Am J Roentgenol* 2007;189(3):616-623.
- Andersson I, Ikeda DM, Zackrisson S, et al. 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* 2008;18(12):2817-2825.
- Spangler ML, Zuley ML, Sumkin JH, et al. Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: a comparison. *AJR Am J Roentgenol* 2011;196(2):320-324.
- Gur D, Abrams GS, Chough DM, et al. Digital breast tomosynthesis: observer performance study. *AJR Am J Roentgenol* 2009;193(2):586-591.
- Gennaro G, Toledano A, di Maggio C, et al. Digital breast tomosynthesis versus

- digital mammography: a clinical performance study. *Eur Radiol* 2010;20(7):1545–1553.
17. Mammography Quality Standards Act of 1992. Public Law 102-539. As amended by the Mammography Quality Standards Reauthorization Act of 1998, Pub. L. No. 105-248, Title 42, Subchapter II, Part F, Subpart 3, § 354 (42 USC 263b), certification of mammography facilities.
 18. D'Orsi CJ, Bassett LW, Berg WA, et al. *Breast Imaging Reporting and Data System, BI-RADS: Mammography*. 4th ed. Reston, Va: American College of Radiology, 2003.
 19. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jack-knife method. *Invest Radiol* 1992;27(9):723–731.
 20. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol* 2008;15(5):647–661.
 21. Gur D, Bandos AI, Rockette HE, et al. Is an ROC-type response truly always better than a binary response in observer performance studies? *Acad Radiol* 2010;17(5):639–645.
 22. US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2009;151(10):716–726, W-236.
 23. Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Ann Intern Med* 2011;155(8):481–492.